EMPIRICAL BAYES WITH A CHANGING PRIOR

A thesis

submitted in fulfilment

of the requirements for the Degree

of

Doctor of Philosophy in Mathematics

in the

University of Canterbury

by

Michael Kelly Mara

University of Canterbury

January, 1979

To Tricia

CONTENTS

CHAPTER PAGE

ABSTRACT

This thesis examines a modified empirical Bayes decision problem in which the a priori distribution is assumed to change with each successive component problem. Firstly, the usual empirical Bayes problem and associated rates of convergence to optimality is considered and rates are given for certain parametric examples. General results on asymptotic optimality for the modified problem are given. The problem is then examined for location parameter families in which the a priori mean is assumed to change in a linear fashion. Asymptotically optimal estimators are given for both the two action problem and the estimation problem under squared error loss. Generalised convergence rates are developed for the exponential family of densities. Some examples of parametric a priori distributions are also considered.

The problem of the selection of the 'best' of several populations is studied in the modified empirical Bayes framework. Asymptotically optimal procedures for this problem, under a linear loss structure, are developed. The modified empirical Bayes method is applied to a system reliability model. Finally the modified problem is extended to examine the case where the a priori means are random variables, generated by a martingale process. Asymptotically optimal estimators, with associated rates of convergence, are established here under the assumption that the a priori distributions are members of a parametric family.

CHAPTER I

INTRODUCTION

I. GENERAL INTRODUCTION AND SUMMARY OF CONTENTS

Consider the following problem. A random variable X, assuming values in a set $\mathcal{X}$ is distributed according to a distribution function in the class $\{P_\theta(x): \theta \in \Omega\}$. For each value $\theta \in \Omega$, $P_\theta(x)$ is completely specified. If an observation $X = x$ is taken, we are required to make a decision $t(x)$ about the unknown value of θ. For example this may be the calculation of a point estimate of θ or a choice between hypothetical values of θ. Dependence of the decision on the observation x is indicated by the use of $t(x)$, a function which maps the set $\mathcal{X}$ into a set A of possible actions (or decisions). In any given decision problem, there will be many possible decision functions and the problem is to select one which has acceptable properties.

In the Bayesian approach to the problem, the parameter θ is regarded as a random variable assuming values in Ω according to the distribution function $G(\theta)$. However in order to apply the Bayesian method, randomness of the parameter θ is not sufficient. The a priori distribution G must be known.

In the empirical Bayes approach existence of G is postulated, but it is not assumed to be known. It is supposed that certain previous data is available, suitable for the estimation of G or for the estimation of the Bayes procedure

$t_G(x)$. In particular we have the sequence $X_1, X_2, \ldots X_n$ of observed values, each X_i is drawn independently according to the density $f(\cdot|\theta)$. The object is then to obtain a sequence of decision rules $t_n(x) = t_n(X_1, \ldots, X_n; x)$ which converge, according to suitable criteria,to the optimal Bayes rule.

This general approach has been established in a paper by Robbins (1955) who has developed empirical Bayes procedures for certain estimation problems which are asymptotically optimal in the sense that the expected risk for the nth decision problem converges to the optimal Bayes risk. Since then the problem has received a great deal of attention and the results of Robbins have been extended by Johns (1957), Samuel (1963), Robbins (1963, 1964), Deely (1965) to mention only a few.

In parts II and III of this chapter, the Bayesian and empirical Bayesian decision models will be formally set up. The remainder of the chapter gives some optimality results due to Robbins (1964).

The usefulness of empirical Bayes decision procedures in any practical statistical applications will depend on the rapidity with which the risks, incurred in successive decision problems, approach the optimal Bayes risk. This question of rate of convergence was first raised in Robbins (1963). Johns and Van Ryzin (1971, 1972) have provided an excellent study of generalised convergence rates for the hypothesis testing problem under a linear loss structure, with the assumption that $f(x|\theta)$ is a member of the exponential

family of densities. Lin (1971, 1975) has extended the analysis to the problem of squared error loss estimation of θ. These results are reported in Chapter II and some analysis is performed on their small sample properties. If a parametric assumption is made on the a priori distribution then the empirical Bayes problem is made easier. Rates of convergence and small sample properties are generally much better in this case than for the case of nonparametric G. Several examples of this type are also studied in Chapter II.

One of the principal assumptions in the empirical Bayesian method is that the component problems are identical with the a priori distribution $G(\theta)$ being fixed for every successive decision problem. Suppose however that the a priori distribution is itself subject to change. That is, at stage i, the a priori distribution is $G_i(\theta)$. This change must be a structured one because if not, the observations $X_1, \ldots, X_n$ taken on $f(x|\theta_1), \ldots, f(x|\theta_n)$, with θ_i distributed according to G_i, $i = 1, \ldots, n$, would not be useful for determining the value of the 'present' parameter θ_{n+1}.

In Chapter III the problem of changing G is examined and some general optimality results are derived. It is then assumed that $f(x|\theta)$ and $g(\theta) = G'(\theta)$ are location parameter densities and that the change in the a priori distribution takes the form of a linear change in the mean. Asymptotically optimal procedures are then established for both the two action problem under linear loss and the squared error loss estimation problem. Rates of convergence to optimality for

exponential family densities are given. Examples for which a parametric form is assumed for the $\{G_i\}$ are also given and, as with the usual empirical Bayes case, it is found that these rates are better than for the nonparametric case.

The problem of selecting the 'best' of several populations, (sometimes referred to as the vendor-ranking problem), is important to several applications. For example, a manufacturer wanting to buy quantities of a component from k possible suppliers will want to determine which supplier produces the component with the longest mean lifetime. Several authors have studied this problem using empirical Bayesian methods including Deely (1965), Van Ryzin (1970), Van Ryzin and Susarla (1977). In Chapter IV the study of changing G is continued. In fact it is assumed that the a priori distribution of each of the populations changes in the manner described in Chapter III. Asymptotically optimal procedures are given for selecting best population under a linear loss structure and, for the normal-normal case, under zero-one loss.

The modified empirical Bayes model is applied to a model of system reliability in Chapter V. Two examples are considered: one where the system is subject to wearout failure and also one in which a random failure may occur.

The general model discussed above is extended in Chapter VI to cover the case where the change in the a priori mean is a stochastic one. It is assumed that G_n is a member of a parametric family and that its mean, λ_n, follows a martingale process. Using results of martingale theory, empirical Bayes procedures are given and shown to be asymptotically optimal,

according to a modified definition. Examples of this type are then considered. This thesis concludes with a short discussion.

II THE BAYES APPROACH

The basic structure of the statistical decision problem we consider is as follows:

(a) A parameter space Ω, with generic element θ. The parameter value θ or "state of nature" will generally be unknown.

(b) An action space A, with generic element a.

(c) A loss function $L(a,\theta) \geq 0$, which represents the loss we incur upon choosing action a when the value of the parameter is θ.

(d) An a priori distribution G on Ω.

(e) An observable random variable X, assuming values in a space $\mathcal{X}$ on which a σ-finite measure μ is defined. When the parameter is θ, X will have the specified probability density function f_θ with respect to μ.

The general problem is to choose a decision function t, defined on $\mathcal{X}$ and assuming values in A, such that when we have observed $X = x$, we will take action $t(x)$ and thereby incur the loss $L(t(x), \theta)$.

For any decision function t, the expected loss incurred when the parameter is $\theta \in \Omega$, is

$$R(t,\theta) = \int_{\mathcal{X}} L(t(x),\theta) f_\theta(x) d\mu(x) \quad . \tag{1.1}$$

Since it has been assumed that θ is a random variable with a priori distribution G , $R(t,\theta)$ will be a random function and the true expected loss will be the Bayes risk, defined by

$$R(t,G) = \int_{\Omega} R(t,\theta) dG(\theta)$$

$$= \int_{\Omega} \int_{\mathcal{X}} L(t(x),\theta) f_\theta(x) d\mu(x) dG(\theta) \quad . \tag{1.2}$$

Provided $L(t(x),\theta) f_\theta(x)$ is integrable on the product space of $\mathcal{X}$ and Ω, use of Fubini's theorem will give

$$R(t,G) = \int_{\mathcal{X}} [\int_{\Omega} L(t(x),\theta) f_\theta(x) dG(\theta)] d\mu(x) \quad . \tag{1.3}$$

Setting

$$\phi_G(a,x) = \int_{\Omega} L(a,\theta) f_\theta(x) dG(\theta) \tag{1.4}$$

for each $a \in A$, we have

$$R(t,G) = \int_{\mathcal{X}} \phi_G(a,x) d\mu(x) \quad . \tag{1.5}$$

If a decision procedure t_G has the property

$$R(t_G,G) \leqslant R(t,G) \tag{1.6}$$

for all decision procedures t, then t_G is defined as the Bayes decision procedure with respect to the a priori distribution G.

Assume now that there exists a decision rule t_G with the property that, for almost all $x \in \mathcal{X}$

$$\phi_G(t_G(x),x) = \min_{a\in A} \phi_G(a,x). \tag{1.7}$$

Thus, for any other decision function t, for almost all $x \in \mathcal{X}$

$$\phi_G(t_G(x),x) \leqslant \phi_G(t(x),x). \tag{1.8}$$

Condition (1.8) now guarantees (1.6). Consequently we may define

$$R(G) = R(t_G,G) = \int_{\mathcal{X}} \phi_G(t_G(x),x)d\mu(x) \tag{1.9}$$

and it follows that

$$R(G) = \min_{t} R(t,G) . \tag{1.10}$$

R(G) is called the Bayes envelope functional. When G is a known distribution we may use t_G to attain the minimum possible Bayes risk, R(G).

Application of this 'Bayesian' method demands that G is a known distribution function. The empirical Bayes approach is concerned with the problem where existence of G is postulated, but that it is initially unknown to the decision maker. Thus the Bayes decision procedure t_G is not available to us, but if we are in a repetitive decision situation it is possible to

use observed data to estimate G and to base the next decision on this estimate. If the data and the sequential decision procedure are of a suitable form, then it is possible under certain conditions to provide, at each successive stage of observation, estimates of G which converge to G and for which the decision procedures corresponding to these estimates converge to t_G.

III THE EMPIRICAL BAYES APPROACH: MODEL AND DEFINITIONS

The empirical Bayes approach, introduced by Robbins (1955) tries to overcome this difficulty of lack of knowledge of G. The basic assumption is that the experiment at hand was preceded by a series of comparable experiments. Generally speaking there are three possible approaches:

(i) Estimate G, say by $\hat{G}$, and use $t_{\hat{G}}$ as an approximation for t_G.

(ii) Estimate $\phi_G(\cdot,x)$ (defined in (1.4)), say by $\hat{\phi}_G(\cdot,x)$ and use the decision rule which minimises $\hat{\phi}_G(\cdot,x)$.

(iii) Estimate $t_G(x)$ directly.

Specifically, suppose $(\theta_1,X_1),(\theta_2,X_2), \ldots, (\theta_n,X_n)$ is a sequence of pairs of random variables, each pair being independent of all other pairs. The θ_n are identically distributed according to the a priori distribution G and the conditional density function of X_n given $\theta = \theta$ is

specified by $f_\theta(\cdot)$. At stage n+1, i.e. when the decision about the value of θ_{n+1} has to be made, the "prior" observations $X_1 = x_1, \ldots, X_n = x_n$, as well as the "present" observation $X_{n+1} = x_{n+1}$, are available. (The values $\Theta_1 = \theta_1, \ldots, \Theta_n = \theta_n$ will in most cases remain unknown). Thus, any decision about θ_{n+1} may be made by a function of x_{n+1} whose form depends on $x_1, \ldots, x_n$.

Formally, let

$$t_n(\cdot) = t_n(x_1, \ldots, x_n; \cdot) \tag{1.11}$$

be a mapping from $\mathcal{X}$ into A such that upon taking action $t_n(x)$ in A, the loss incurred is $L(t_n(x),\theta)$. We then define an empirical Bayes decision procedure to be a sequence $T = \{t_n\}$ of functions of the form (1.11). As in the previous section we define the risk, or expected loss for a given sequence $(x_1, \ldots, x_n)$ as

$$r_n(T,G) = \int_{\mathcal{X}} \phi_G(t_n(x),x)\,d\mu(x) \tag{1.12}$$

where ϕ_G is defined in (1.4). The overall expected loss, or Bayes risk of the sequence T relative to G, is

$$R_n(T,G) = \int_{\mathcal{X}} E[\phi_G(t_n(x);x)]\,d\mu(x) \tag{1.13}$$

where E denotes expectation with respect to the independent variables $X_1, \ldots, X_n$, given by

$$E[\phi_G(t_n(x),x)]$$

$$= \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} \phi_G(t_n(x),x)\ f_G(x_1)dx_1 \ \ldots\ f_G(x_n)dx_n \qquad (1.14)$$

where

$$f_G(x_i) = \int_{\Omega} f_\theta(x_i)dG(\theta), \qquad i = 1,2, \ \ldots\ , n, \qquad (1.15)$$

is the marginal probability density function for each of the random variables $X_1, \ \ldots\ , X_n$. From (1.6) and (1.13) it follows that, for all n

$$R_n(T,G) \geqslant R(G) \ . \qquad (1.16)$$

Intuitively, the closeness of the quantity $R_n(T,G)$ to $R(G)$ would seem to be a good measure of the effectiveness of an empirical Bayes decision procedure T and a desirable property of T would therefore be the convergence of $R_n(T,G)$ to $R(G)$.

Definition

The sequence of decision procedures $T = \{t_n\}$ is said to be asymptotically optimal (a.o.) relative to G if

$$\lim_{n\to\infty} R_n(T,G) = R(G) \ . \qquad (1.17)$$

Most of the literature on empirical Bayes has been concerned with the possibility of constructing empirical Bayes rules which are a.o. for all priors G in some class $\mathcal{G}$.

IV OPTIMALITY RESULTS

The general results on asymptotic optimality in this section are due to Robbins (1964), generalising those of Robbins (1955) and Samuel (1963). Using (1.9) and (1.13) we first note that

$$0 \leqslant R_n(T,G) - R(G) = \int_{\mathcal{X}} \{E[\phi_G(t_n(x),x)] - \phi_G(t_G(x),x)\} d\mu(x) . \tag{1.18}$$

We also make the rather restrictive assumption

$$L(\theta) = \sup_{a \in A} L(\theta,a) < \infty, \quad \forall\ \theta \in \Omega . \tag{1.19}$$

Lemma 1.1

The sequence $T = \{t_n\}$ is asymptotically optimal if (1.19) holds and if

$$p \lim_{n\to\infty} \phi_G(t_n(x),x) = \phi_G(t_G(x),x) \qquad \text{a.e.}[\mu]x , \tag{1.20}$$

[where p lim denotes convergence in probability].

Proof

Define $H(x) = \int_{\Omega} L(\theta) f_\theta(x) dG(\theta)$.

From (1.4) we have a.e. [μ]

$$\phi_G(a,x) \leqslant H(x) \quad , \quad \forall\ a \in A .$$

Thus, for any $T = \{t_n\}$,

$$\phi_G(t_n(x),x) \leqslant H(x) \quad , \quad \forall n \quad , \quad \text{a.e. } [\mu]$$

and by assumption

$$\int_{\mathcal{X}} H(x)d\mu(x) = \int_{\Omega} L(\theta)dG(\theta) < \infty \quad . \tag{1.21}$$

Hence $H(x) < \infty$ a.e.$[\mu]$ and the Helly-Bray theorem gives

$$\lim_{n\to\infty} E[\phi_G(t_n(x),x)] = \phi_G(t_G(x),x) \qquad \text{a.e.}[\mu] \quad . \tag{1.22}$$

Furthermore

$$0 \leqslant E[\phi_G(t_n(x),x)] - \phi_G(t_G(x),x) \leqslant H(x)$$

and it follows, from the Lebesgue Dominated Convergence theorem and (1.22) that

$$\lim_{n\to\infty}\{R_n(T,G) - R(G)\} = 0$$

i.e. $T = \{t_n\}$ is asymptotically optimal.

The following method has been suggested by Robbins (1964) as a means to find an a.o. empirical Bayes rule.

Suppose a_o is any arbitrary element of A and define

$$\Delta_G(a,x) = \int_{\Omega} [L(a,\theta) - L(a_o,\theta)]f_\theta(x)dG(\theta) \quad ,$$

$$L_o(x) = \int_{\Omega} L(a_o,\theta)f_\theta(x)dG(\theta) \quad . \tag{1.23}$$

Thus we have from (1.4)

$$\phi_G(a,x) = L_o(x) + \Delta_G(a,x). \tag{1.24}$$

Provided $L_o(x) < \infty$ (a condition guaranteed by (1.19)) then in order to satisfy (1.20) it is sufficient to find a sequence, $T = \{t_n\}$, of decision procedures which satisfies

$$p \lim_{n\to\infty} \Delta_G(t_n(x),x) = \Delta_G(t_G(x),x) \qquad \text{a.e.}[\mu] \ . \tag{1.25}$$

Since by assumption, G is not known it follows that neither $t_G(x)$ nor $\Delta_G(\cdot,x)$ are known. Hence, before finding a sequence of decision procedures $\{t_n\}$ which is asymptotically optimal, it is necessary to find a sequence $\Delta_n(a,x)$ which converges in probability to $\Delta_G(a,x)$ for all $a \in A$, for almost all $x \in \mathcal{X}$.

Theorem 1.2

Let G be such that (1.19) holds. Let $\Delta_n(a,x) = \Delta_n(x_1, \ldots, x_n; a, x)$ be a sequence of functions satisfying the condition

$$p \lim_{n\to\infty} \sup_{a\in A} |\Delta_n(a,x) - \Delta_G(a,x)| = 0 \qquad \text{a.e.}[\mu]x \ . \tag{1.26}$$

Then the sequence of the decision functions $T = \{t_n(x)\}$, defined as

$$t_n(x) = t_n(x_1, \ldots, x_n; x) = a^* \in A$$

such that

$$\Delta_n(a^*,x) \leqslant \inf_{a\in A} \Delta_n(a,x) + \varepsilon_n$$

for any sequence ε_n tending to zero, is asymptotically optimal relative to G.

Proof

By (1.24)

$$0 \leqslant \phi_G(t_n(x),x) - \phi_G(t_G(x),x)$$

$$= \Delta_G(t_n(x), x) - \Delta_G(t_G(x),x)$$

$$= [\Delta_G(t_n(x),x) - \Delta_n(t_n(x),x)] + [\Delta_n(t_n(x),x)$$

$$- \Delta_n(t_G(x),x)] + [\Delta_n(t_G(x),x) - \Delta_G(t_G(x),x)] \quad . \qquad (1.27)$$

Condition (1.26) guarantees that for any $\varepsilon > 0$ there exists an N such that $n \geqslant N$ implies

$$|\Delta_G(t_n(x),x) - \Delta_n(t_n(x),x)| < \varepsilon$$

$$|\Delta_n(t_G(x),x) - \Delta_G(t_G(x),x)| < \varepsilon$$

with probability greater than $1-\varepsilon$. Hence by the definition of $t_n(x)$, the right hand side of (1.27) becomes

$$< \varepsilon + \varepsilon_n + \varepsilon$$

with probability as near to 1 as we like for n sufficiently large.

$$\text{i.e.} \quad p \lim_{n\to\infty} \phi_G(t_n(x),x) = \phi_G(t_G(x),x) \qquad \text{a.e.}[\mu]$$

and thus by Lemma 1.1, $T = \{t_n\}$ is asymptotically optimal with respect to G.

Unfortunately the result does not give a method for determining the sequence $\{\Delta_n\}$, so that in its most general form this theorem is not useful for applications. However, when this result is applied to finite action problems, and in particular to hypothesis testing problems, its simpler formulation renders it useful for a range of common distributions.

Corollary 1

Let $A = \{a_o, \ldots, a_m\}$ be a finite set and let G be such that

$$\int_\Omega L(a_i,\theta)dG(\theta) < \infty \quad , \quad i\varepsilon\{0, \ldots, m\} \tag{1.28}$$

and let $\Delta_{i,n}(x) = \Delta_{i,n}(x_1, \ldots, x_n; x)$ for $i = 1, \ldots, m$ and $n = 1,2, \ldots$ be such that for a.e. $[\mu]x$,

$$p \lim_{n\to\infty} \Delta_{i,n}(x) = \int_\Omega [L(a_i,\theta) - L(a_o,\theta)]f_\theta(x)dG(\theta) . \tag{1.29}$$

Set $\Delta_{o,n}(x) = 0$ and define $t_n(x) = a_k$ where k is any integer $0 \leqslant k \leqslant m$ such that

$$\Delta_{k,n}(x) = \min\{0,\Delta_{1,n}(x), \ldots, \Delta_{m,n}(x)\} \quad .$$

Then $T = \{t_n\}$ is asymptotically optimal relative to G.

Corollary 2 (Hypothesis Testing)

Let $A = \{a_o,a_1\}$ and let G be such that

$$\int_\Omega L(a_i,\theta)dG(\theta) < \infty \quad , \quad i = 0,1,$$

and let

$\Delta_n(x) = \Delta_n(x_1, \ldots, x_n; x)$ be such that for a.e.[μ]x

$$\underset{n\to\infty}{p\lim}\ \Delta_n(x) = \Delta_G(x)$$

$$= \int_{\Omega} [L(a_1,\theta) - L(a_o,\theta)]f_\theta(x)dG(\theta) \ . \qquad (1.30)$$

Define $T = \{t_n\}$ as

$$t_n(x) = \begin{cases} a_o \ , & \text{if } \Delta_n(x) \geqslant 0 \\ a_1 \ , & \text{if } \Delta_n(x) < 0 \end{cases} \ . \qquad (1.31)$$

Then $T = \{t_n\}$ is asymptotically optimal relative to G.

V EXAMPLES OF EMPIRICAL BAYES RULES

(1) Hypothesis Testing

Consider the example of testing the hypotheses

$$H_o: \theta < \theta_o \quad \text{vs} \quad H_1: \theta > \theta_o$$

for some θ_o. Based on the observation X we shall make one of two decisions. i.e. $A = \{a_o, a_1\}$, where a_o is equivalent to accepting H_o and a_1 rejects H_o. Following Johns (1957) we use the linear loss function

$$L(a_i,\theta) = \begin{cases} \max\{0, \theta-\theta_o\} \ , & i = 0 \\ -\min\{0, \theta-\theta_o\} \ , & i = 1 \end{cases} \ . \qquad (1.32)$$

This yields

$$L(a_1,\theta) - L(a_o,\theta) = \theta_o - \theta \qquad , \qquad \forall\ \theta \in \Omega$$

which, from (1.23), gives

$$\begin{aligned}\Delta_G(x) &= \Delta_G(a_1,x) \\ &= \int_\Omega [L(a_1,\theta) - L(a_o,\theta)] f_\theta(x)\, dG(\theta) \\ &= \int_\Omega (\theta_o - \theta) f_\theta(x)\, dG(\theta) \quad . \end{aligned} \tag{1.33}$$

Consider the family of Poisson distributions with parameter θ, $\theta > 0$.

$$f_\theta(x) = \frac{\theta^x e^{-\theta}}{x!} \ , \ x = 0,1,2, \ \ldots \tag{1.34}$$

Hence,

$$\begin{aligned}\Delta_G(x) &= \int_o^\infty (\theta_o - \theta)\, \frac{\theta^x e^{-\theta}}{x!}\, dG(\theta) \\ &= \theta_o f_G(x) - (x+1) f_G(x+1) \ , \end{aligned} \tag{1.35}$$

where

$$f_G(x) = \int_o^\infty f_\theta(x)\, dG(\theta)$$

is the unconditional density of X.

Corollary 2 of Theorem 1.2 requires a sequence $\Delta_n(x)$ converging in probability to $\Delta_G(x)$. From (1.34) we see that what is needed is an estimate of $f_G(x)$. The natural estimator is just the empirical density function

$$f_n(x) = \frac{\#\text{ of } x_1, \ldots, x_n = x}{n} \quad . \tag{1.36}$$

The Weak Law of Large Numbers gives

$$f_n(x) \xrightarrow{p} f_G(x) \qquad \text{a.e.}[\mu].$$

Hence, forming

$$\Delta_n(x) = \theta_o f_n(x) - (x+1)\, f_n(x+1)$$

we have satisfied the conditions of the corollary for all distributions G with a finite first moment. The empirical Bayes procedure is now

$$t_n(x) = \begin{cases} a_o \ , & \text{if } \Delta_n(x) \geqslant 0 \\ a_1 \ , & \text{if } \Delta_n(x) < 0 \end{cases} ,$$

and this procedure is asymptotically optimal.

Other examples of this type may be found in Robbins (1963) and Samuel (1963).

(2) Estimation

Suppose $A = \Omega$ and consider the problem of estimating the parameter $\theta \in \Omega$ under the following (squared-error) loss function.

$$L(a,\theta) = (a-\theta)^2 \tag{1.37}$$

Immediately, for any decision procedure t

$$\phi_G(t(x),x) = \int_\Omega (t(x) - \theta)^2 f_\theta(x)\, dG(\theta) \quad . \tag{1.38}$$

It is well known that $\phi_G(t(x),x)$ is minimised over t by the choice of $t_G(x)$ as the posterior mean

$$t_G(x) = \frac{\int_\Omega \theta f_\theta(x) dG(\theta)}{\int_\Omega f_\theta(x) dG(\theta)}, \tag{1.39}$$

the denominator being the unconditional density $f_G(x)$.

As an example of this consider again the Poisson distribution with parameter $\theta > 0$ in (1.33). The Bayes rule is

$$t_G(x) = E[\theta | X=x] = \frac{\int_0^\infty \theta \cdot \frac{\theta^x e^{-\theta}}{x!} dG(\theta)}{f_G(x)}$$

$$= (x+1) \frac{f_G(x+1)}{f_G(x)} \quad . \tag{1.40}$$

Using the empirical density function as before we have

$$t_n(x) = (x+1) \frac{f_n(x+1)}{f_n(x)} \quad . \tag{1.41}$$

We cannot establish asymptotic optimality using Theorem 1.2 because the loss function does not satisfy (1.19). This is a general difficulty in the problem of squared error loss estimation. However, for this case Johns (1957) has shown that $t_n(x)$ is asymptotically optimal for all priors G for which

$$\int_\Omega \theta^2 dG(\theta) < \infty \quad .$$

The following lemma is often of assistance when computing the difference in Bayes risk between the Bayes estimator $t_G(x)$ and an empirical Bayes estimator $t_n(x)$.

Lemma 1.3

$$R(T,G) - R(G) = \int_{\mathcal{X}} E[t_n(x) - t_G(x)]^2 \, f_G(x) d\mu(x) \qquad (1.42)$$

where, as before, E denotes expectation over the n random variables $X_1, \ldots, X_n$.

Proof

$$R(T,G) = \int_{\mathcal{X}} \int_{\Omega} E[t_n(x) - \theta]^2 f_\theta(x) dG(\theta) d\mu(x)$$

$$= \int_{\mathcal{X}} \int_{\Omega} E[t_n(x) - t_G(x)]^2 f_\theta(x) dG(\theta) d\mu(x)$$

$$+ \int_{\mathcal{X}} \int_{\Omega} (t_G(x) - \theta)^2 f_\theta(x) dG(\theta) d\mu(x)$$

$$+ 2 \int_{\mathcal{X}} \int_{\Omega} E[t_n(x) - t_G(x)][t_G(x) - \theta] f_\theta(x) dG(\theta) d\mu(x)$$

By (1.38) the third term on the right hand side is zero and the second term is simply the Bayes risk for the Bayes rule R(G).

From this lemma we see that asymptotic optimality of t_n implies that, as $n \to \infty$

$$t_n(x) \xrightarrow{p} t_G(x) \qquad \text{a.e.}[\mu] \quad .$$

However, in contrast to the hypothesis testing problems, it is not easy to find general conditions to give asymptotic optimality for this estimation problem. In general we need to consider particular estimators to determine asymptotic optimality.

CHAPTER II

RATES OF CONVERGENCE TO OPTIMALITY

I. INTRODUCTION

When considering an empirical Bayes decision rule $T = \{t_n\}$, then clearly asymptotic optimality is the minimum desired goal. However, a very important factor, which must be taken into account when considering the applicability of such rules, is the number of "previous" observations that are necessary in order to achieve some required degree of accuracy. For a given $\varepsilon > 0$ we require an $N = N(\varepsilon)$ such that

$$0 \leqslant R_n(T,G) - R(G) < \varepsilon, \qquad \text{for all } n \geqslant N. \tag{2.1}$$

i.e. we require a rate of convergence to optimality. This subject has received some attention in recent years, significant contributors being Johns and Van Ryzin (1971), (1972) and Lin (1971), (1975).

(1) Discrete Case

(a) Finite Action Problem

Consider the hypothesis testing problem

$$H_0: \ \theta < \theta_0 \quad \text{vs} \quad H_1: \ \theta \geqslant \theta_0$$

under the loss structure (1.32) for the case where the

observations have the conditional probability mass function

$$f_\theta(x) = h(x)\theta^x\beta(\theta) \quad , \qquad x = 0, 1, 2, \ldots \qquad (2.2)$$
$$0 \leqslant \theta < d.$$

The Poisson, geometric and negative binomial distributions are members of this family. We have $A = \{a_0, a_1\}$ and with a priori distribution G and using a decision procedure $t(x) = \Pr(\text{accepting } H_0 \mid X = x)$, the Bayes risk is

$$\begin{aligned} R(t,G) &= \sum_x \{t(x) \int_\Omega [L(a_0,\theta) - L(a_1,\theta)] f_\theta(x) dG(\theta)\} \\ &\quad + \sum_x \int_\Omega L(a_1,\theta) f_\theta(x) dG(\theta) \\ &= \sum_x \alpha(x)t(x) + \int_\Omega L(a_1,\theta) dG(\theta) \quad , \end{aligned} \qquad (2.3)$$

where

$$\begin{aligned} \alpha(x) &= \int_\Omega \theta f_\theta(x) dG(\theta) - \theta_0 \int_\Omega f_\theta(x) dG(\theta) \\ &= \frac{h(x)}{h(x+1)} f(x+1) - \theta_0 f(x) \quad . \end{aligned} \qquad (2.4)$$

[f(x) being the unconditional probability mass function of the observation X].

To ensure finiteness of the Bayes risk we will assume

$$\int_\Omega \theta dG(\theta) < \infty \ .$$

The Bayes procedure for this problem is now

$$t_G(x) = \begin{cases} 1 , & \text{if } \alpha(x) \leqslant 0 \\ 0 , & \text{if } \alpha(x) > 0 \end{cases} . \tag{2.5}$$

As before this rule is known only if G is a known distribution. When G is not known we must construct a sequence of estimates $\alpha_n(x)$ of $\alpha(x)$ using the observations $X_1, \ldots, X_n$. From this we obtain an empirical Bayes procedure analagous to (2.5) as

$$t_n(x) = \begin{cases} 1 , & \text{if } \alpha_n(x) \leqslant 0 \\ 0 , & \text{if } \alpha_n(x) > 0 \end{cases} . \tag{2.6}$$

From (2.4) it is seen that estimation of the marginal density $f(x)$ is sufficient for the estimation of $\alpha(x)$. The natural estimator is again the empirical probability mass function $f_n(x)$ of (1.36), and we form

$$\alpha_n(x) = \alpha_n(x_1, \ldots, x_n; x)$$

$$= \frac{h(x)}{h(x+1)} f_n(x+1) - \theta_0 f_n(x) . \tag{2.7}$$

By the Law of Large Numbers $f_n(x) \xrightarrow{p} f(x)$ a.e. $[\mu]$ and hence $\alpha_n(x) \xrightarrow{p} \alpha(x)$ a.e. $[\mu]$. Thus, corollary 2 to Theorem 1.2 yields asymptotic optimality for the procedure (2.7) provided only that $E[\theta] < \infty$. The following rate result is due to Johns and Van Ryzin (1971).

Define $\eta_n(x) = \mathrm{Var}\{\alpha_n(x)\}$.

Theorem 2.1

Consider the discrete exponential family (2.2) under the loss function (1.32). If for some $0 < \delta < 2$ there exists a $K > 0$ such that

$$\sum_{\{x\in\mathcal{X}:\,\alpha(x)\neq 0\}} |\alpha(x)|^{1-\delta}\{\sqrt{n}\,\eta_n(x)\}^{\delta} < K < \infty \tag{2.8}$$

then with $T = \{t_n\}$ given by (2.6) and (2.7)

$$0 \leqslant R_n(T,G) - R(G) \leqslant Kn^{-\frac{1}{2}\delta} \quad . \tag{2.9}$$

Proof:

From (2.3)

$$0 \leqslant R_n(T,G) - R(G)$$

$$= \sum_{x\in\mathcal{X}} \alpha(x)E[t_n(x)] - \sum_{x\in\mathcal{X}} \alpha(x)t_G(x) \quad .$$

$$= \sum_{x\in\mathcal{X}} \alpha(x)[\Pr(\alpha_n(x) \leqslant 0) - t_G(x)]$$

$$= \sum_{x\in\mathcal{X}} \alpha(x)Q_n(x) \quad ,$$

where

$$Q_n(x) = \begin{cases} \Pr(\alpha_n(x) > 0), & \text{if } \alpha(x) \leqslant 0 \\ \Pr(\alpha_n(x) \leqslant 0), & \text{if } \alpha(x) > 0 \end{cases}$$

$$\leqslant \Pr(|\alpha_n(x) - \alpha(x)| \geqslant |\alpha(x)|)$$

$$= \Pr(|\alpha_n(x) - \alpha(x)|^{\delta} \geqslant |\alpha(x)|^{\delta})$$

$$\leqslant \frac{E|\alpha_n(x) - \alpha(x)|^{\delta}}{|\alpha(x)|^{\delta}} \quad ,$$

by the Markov inequality. This gives the desired result.

This theorem as it stands is basically unintuitive and the following corollaries give simpler conditions for asymptotic optimality.

Corollary 1:

Consider the family (2.2). If for some $0 < \delta < 2$ there exist constants C_0, C_1 such that

$$\text{(i)} \quad \max\left\{\frac{h(x)}{h(x+1)} \ , \quad \frac{f(x+1)}{f(x)}\right\} < C_0 < \infty$$

$$\text{(ii)} \quad \left|\frac{h(x)}{h(x+1)} \cdot \frac{f(x+1)}{f(x)} - \theta_0\right| > C_1 > 0$$

for all sufficiently large x, and

$$\text{(iii)} \quad \sum_{x \in \mathcal{X}} [f(x)]^{1-\frac{1}{2}\delta} < \infty$$

then there exists a $K > 0$ such that

$$0 \leqslant R_n(T,G) - R(G) \leqslant Kn^{-\frac{1}{2}\delta} \quad .$$

Corollary 2:

If there exists a constant C_0 such that

$$\text{(i)} \quad \frac{f(x)}{f(x+1)} < C_0 < \infty$$

(ii) $\frac{h(x)}{h(x+1)} \to \infty$, as $x \to \infty$

(iii) $\sum_{x \in \mathcal{X}} \frac{h(x)}{h(x+1)} [f(x+1)]^{1-\frac{1}{2}\delta} < \infty$

then, for the family (2.2), there exists a $K > 0$ such that (2.9) holds.

[$f(x)$ is the unconditional p.d.f. of X]

Example

Suppose that $X_1, \ldots, X_n$ is a sequence of independent and identically distributed random variables, each with the conditional probability mass function of (2.2) with $h(x) \equiv 1$ and $\beta(\theta) = 1-\theta$. i.e. we have the geometric mass function with parameter θ.

$$f_\theta(x) = \theta^x(1-\theta) \qquad x = 0,1,2, \ldots; \ 0 \leqslant \theta < 1 . \quad (2.10)$$

In addition suppose that the prior distribution G has the following density

$$G'(\theta) = g(\theta) = (a+1)(1-\theta)^a, \qquad 0 \leqslant \lambda < 1; \ a > -1 . \quad (2.11)$$

This yields the unconditional

$$f(x) = (a+1) \int_0^1 \theta^x (1-\theta)^{a+1} d\theta$$

$$= \frac{(a+1)\Gamma(x+1)\Gamma(a+2)}{\Gamma(a+x+3)} , \quad (2.12)$$

where Γ is the gamma function.

Since

$$\frac{f(x+1)}{f(x)} = \frac{x+1}{a+x+3}$$

we have that on choosing $C_0 = 1$ we can satisfy condition (i) of Corollary 1 to Theorem 2.1. The other two conditions are satisfied for

$$x > \frac{\theta_0(a+3)}{1-\theta_0}$$

and provided

$$\sum_{x=1}^{\infty} x^{-\frac{(a+2)(2-\delta)}{2}} < \infty \quad .$$

These conditions can be satisfied by restricting δ to

$$\delta < \frac{2(1+a)}{2+a} \quad .$$

The corollary now guarantees that for a given value of a we have a rate of convergence arbitrarily close to $0\left(n^{-\frac{a+1}{a+2}}\right)$.

The practical effectiveness of this result must now be determined by numerical example. This will therefore involve evaluation of the constant K in (2.9). From (2.4) we have

$$\alpha(x) = \left(\frac{x+1}{a+x+3} - \theta_0\right) f(x)$$

and therefore

$$\alpha(x) > 0 \quad \text{for} \quad x > \frac{\theta_0(a+3) - 1}{1 - \theta_0} = u, \text{ say.}$$

Also, from (2.7)

$$\eta_n^2(x) \leqslant \frac{1}{n}\{f(x+1) + \theta_0^2 f(x)\}$$

$$= \frac{1}{n}\left\{\frac{x+1}{a+x+3} + \theta_0^2\right\} f(x) \quad .$$

From Theorem 2.1 we therefore need to bound

$$\sum_{x=0}^{[u]} \left(\theta_0 - \frac{x+1}{x+x+3}\right)^{1-\delta} \left(\frac{x+1}{a+x+3} + \theta_0^2\right)^{\frac{1}{2}\delta} \{f(x)\}^{1-\frac{1}{2}\delta}$$

$$+ \sum_{x=[u]+1}^{\infty} \left(\frac{x+1}{a+x+3} - \theta_0\right)^{1-\delta} \left(\frac{x+1}{a+x+3} + \theta_0^2\right)^{\frac{1}{2}\delta} \{f(x)\}^{1-\frac{1}{2}\delta} \qquad (2.13)$$

In particular consider the present problem with $a = 1$, $\theta_0 = 0.5$. This gives $u = 2$ and

$$f(x) = \frac{4\Gamma(x+1)}{\Gamma(x+4)} = \frac{4}{(x+1)(x+2)(x+3)}$$

$$< \frac{4}{(x+1)^3}, \qquad x = 0,1,2, \ldots$$

Choose $\delta = 1.33 < 2(1+a)/2+a) = 4/3$. The value of the first sum in (2.13) is 0.835. Using, for $x > 3$,

$$\frac{x+1}{x+4} - \frac{1}{2} \leqslant \frac{1}{14} \quad \text{and} \quad \frac{x+1}{x+4} + \frac{1}{4} \leqslant \frac{5}{4}$$

it will now be necessary to find a bound for $\sum_{x=3}^{\infty} \left\{\frac{4}{(x+1)^3}\right\}^{1-\frac{1}{2}\delta}$.

This is done using the integral bound. Thus we obtain

$$\sum_{x=0}^{\infty} |\alpha(x)|^{1-\delta}\{\sqrt{n}\eta_n(x)\}^{\frac{1}{2}\delta} \leqslant 0.835$$

$$+ \left(\frac{1}{14}\right)^{1-\delta} 4^{1-\frac{1}{2}\delta}\left(\frac{5}{4}\right)^{\frac{1}{2}\delta} \sum_{x=3}^{\infty} x^{-3(1-\frac{1}{2}\delta)}$$

$$\leqslant 128.48 \quad .$$

Thus the theorem now gives

$$R_n(T,G) - R(G) \leqslant 128.48\ n^{-0.6665} \quad .$$

Remarks

The above example establishes a result of the form (2.1) and we note that for $\varepsilon = 0.1$ say, $N = 46177$. Whether or not this represents a reasonable figure will depend upon the type of experiment. It is, however, necessary to note that although Theorem 2.1 and its corollaries give a general rate of convergence result, the actual convergence rate is not determined until the a priori family is specified. That is, when one knows that the prior is a member of a given parametric family, then it is possible to further determine the rate of convergence to optimality.

Further to this, the determination of the constant K in (2.8) requires the exact specification of the value of a - i.e. the exact prior distribution. This, however, means that the problem is placed into the pure Bayesian framework.

The ideal is to find empirical Bayes rules $\{t_n\}$ which guarantee (2.1) for a sufficiently wide class of a priori distributions.

(b) Estimation

Consider now the problem of point estimation of a parameter $\theta \in \Omega$ under the squared error loss (1.37). As for the two-action case, we consider the discrete exponential family (2.2). The Bayes estimator is the posterior mean $E[\theta|X=x]$ given by (1.39), which for this case is

$$t_G(x) = \frac{h(x)}{h(x+1)} \cdot \frac{f(x+1)}{f(x)} \quad . \tag{2.14}$$

With G unknown we have that using the empirical Bayes rule $T = \{t_n\}$, Lemma 1.3 gives the excess risk as

$$R_n(T,G) - R(G) = \sum_{x=0}^{\infty} E[t_n(x) - t_G(x)]^2 f(x) \quad . \tag{2.15}$$

Observing (2.14) we can find an empirical Bayes estimator by replacing $f(x)$ with the empirical p.m.f. $f_n(x)$ to get

$$t_n(x) = \frac{h(x)}{h(x+1)} \cdot \frac{f_n(x+1)}{f_n(x)} \quad . \tag{2.16}$$

We now have the problem of what to do if $f_n(x) = 0$. To this end suppose $\{\delta_n\}_{n=1}^{\infty}$ is a sequence of positive constants with

$$c_1 n^{-\frac{1}{3}} \leqslant \delta_n \leqslant c_2 n^{-\frac{1}{3}} \quad . \tag{2.17}$$

for some $0 < c_1 \leqslant c_2 < \infty$. The estimator now proposed is

$$t_n^*(x) = \begin{cases} \dfrac{h(x)}{h(x+1)} \dfrac{f_n(x+1)}{f_n(x)} , & \text{if } f_n(x) \geqslant \delta_n \\ \dfrac{h(x)}{h(x+1)} \dfrac{f_n(x+1)}{\delta_n} , & \text{if } f_n(x) < \delta_n \end{cases} . \tag{2.18}$$

The following rate result is due to Lin (1971).

Theorem 2.2

Let $f_\theta(x)$ be given by (2.2) with G an a priori distribution on θ with a finite second order moment and such that $f(x) > 0$ for all x. For the squared error loss problem, if $t_G(x)$ and $t_n^*(x)$ are given by (2.14) and (2.18) respectively and if

$$\text{(i)} \quad \sum_{x=0}^{\infty} \left|\frac{h(x)}{h(x+1)}\right|^2 f(x)f(x+1) < \infty$$

$$\text{(ii)} \quad \sum_{x=0}^{\infty} \left|\frac{h(x)}{h(x+1)}\right|^2 f^2(x+1) < \infty$$

$$\text{(iii)} \quad \sum_{x=0}^{\infty} \left|\frac{h(x)}{h(x+1)}\right|^2 f^2(x+1)/f(x) < \infty$$

and if for some $0 < t \leqslant 1$

$$P(\{x:f(x) < \delta_n\}) < c\delta_n^t$$

with $\{\delta_n\}$ satisfying (2.17), then the sequence of estimators $T = \{t_n\}$ is asymptotically optimal, and moreover

$$R_n(T,G) - R(G) = 0(n^{-t/3}) . \tag{2.19}$$

As before the above conditions are difficult to determine

intuitively and we use an example to illustrate the result.

Example:

Suppose $f_\theta(x)$ is the Poisson mass function (1.34) and that the prior G is such that

$$g(\theta) = G'(\theta) = ae^{-a\theta}, \quad \theta > 0, \quad 0 < a < \infty .$$

In this case $\frac{h(x)}{h(x+1)} = x+1$ and

$$f(x) = \frac{a}{(a+1)^{x+1}}$$

Since $\sum_{x=1}^{\infty} x^2 b^{-x} = b(b+1)(b-1)^{-3}$ for $b > 1$ we have

$$\sum_{x=0}^{\infty} (x+1)^2 f(x) f(x+1) = a^2(a+1)[(a+1)^2+1][(a+1)^2-1]^{-3}$$

$$\sum_{x=0}^{\infty} (x+1)^2 f^2(x+1) = a^2[(a+1)^2+1][(a+1)^2-1]^{-3}$$

$$\sum_{x=0}^{\infty} (x+1)^2 f^2(x+1)/f(x) = (a+2)[a^2(a+1)]^{-1} .$$

Also since $P(\{x: f(x) < \delta_n\}) \leqslant \frac{\delta_n}{1+a}$, we have satisfied the requirements of the theorem with $t = 1$. i.e. we have a rate of convergence of $0(n^{-\frac{1}{3}})$.

As in the hypothesis testing case the evaluation of the bounding constant in this order of convergence requires knowledge of the actual a priori distribution. Furthermore its value is critical in determining N such that (2.1) holds. As an example when $\varepsilon = 0.1$ and $a = 1$, N is as large as 6×10^9.

(2) Continuous Case

(a) A Finite Action Problem

The model for the continuous case two-action problem is similar to that for the discrete case. The observations are assumed to have a density function of the exponential type

$$f_\theta(x) = \begin{cases} e^{-\theta x}\,\beta(\theta)h(x) & , \quad x > a,\ \lambda \text{ in some interval} \\ 0 & , \quad \text{otherwise} \end{cases} \tag{2.20}$$

where a may be finite or infinite and $h(x) > 0$ for $x > a$. The hypothesis $H_0: \theta \leqslant \theta_0$ is to be tested against $H_1: \theta > \theta_0$ under the piecewise linear loss (1.32).

Analagous to the discrete case, the risk of the decision procedure t under the a priori distribution G is

$$R(t,G) = \int_{\mathcal{X}} \alpha(x)t(x)dx + \int_{\Omega} L(a_1,\theta)dG(\theta)$$

where

$$\alpha(x) = \int_{\Omega} \theta f_\theta(x)dG(\theta) - \theta_0 f(x) \quad . \tag{2.21}$$

The Bayes procedure is again given by (2.5) and an empirical Bayes procedure based on $X_1, \ldots, X_n$ is constructed upon estimating $\alpha(x)$ by $\alpha_n(x) = \alpha_n(x_1, \ldots, x_n; x)$, for each x, and using (2.6).

For this exponential family

$$f(x) = h(x) \int_{\Omega} \beta(\theta)e^{-\theta x}dG(\theta) \quad . \tag{2.22}$$

By Theorem 2.9 of Lehman (1959) the above integral is infinitely differentiable and its derivative may be evaluated under the integral sign. Thus differentiability of $f(x)$ will be guaranteed by the differentiability of $h(x)$. Assuming then, the existence of $h^{(1)}(x)$ we may write

$$\int_{\Omega} \theta f_{\theta}(x)\,dG(\theta) = v(x)f(x) - f^{(1)}(x) \tag{2.23}$$

where $v(x) = \dfrac{h^{(1)}(x)}{h(x)}$.

This gives

$$\alpha(x) = (v(x) - \theta_0)f(x) - f^{(1)}(x) \quad , \tag{2.24}$$

so that finding an estimate $\alpha_n(x)$ for $\alpha(x)$ now involves only estimating $f(x)$ and its derivative $f^{(1)}(x)$. The following estimators are based on those given by Parzen (1962) and modified by Johns and Van Ryzin (1972).

Let $f_n(x) = f_n(X_1, \ldots, X_n; x)$ be an estimate of $f(x)$ given by

$$f_n(x) = \frac{1}{na_n} \sum_{j=1}^{n} K_1\left(\frac{X_j - x}{a_n}\right) , \tag{2.25}$$

where $\{a_n\}$ is a sequence satisfying

$$a_n \downarrow 0 \ , \quad na_n^3 \to \infty \quad \text{as} \quad n \to \infty \ ,$$

and $K_1(u)$ is a real-valued measurable function on the real line such that

$$K_1(u) = 0 \text{ if } u \leqslant 0 \text{ or } u \geqslant u_1 > 0$$

$$\int_0^{u_1} K_1(u)du = 1 \tag{2.26}$$

$$\sup_u |K_1(u)| < \infty \quad .$$

To estimate $f^{(1)}(x)$, let $f_n^{(1)}(x) = f_n(X_1, \ldots, X_n; x)$ be given by

$$f_n^{(1)}(x) = \frac{1}{na_n} \sum_{j=1}^{n} \left[\frac{1}{2a_n} K_2\left(\frac{X_j - x}{2a_n}\right) - \frac{1}{a_n} K_2\left(\frac{X_j - x}{a_n}\right) \right] \tag{2.27}$$

where $\{a_n\}$ is as before and $K_2(u)$ is a real valued function for which

$$K_2(u) = 0 \quad \text{if} \quad u \leqslant 0 \quad \text{or} \quad u \geqslant u_2 > 0$$

$$\int_0^{u_2} uK_2(u)du = 1 \tag{2.28}$$

$$\sup_u |K_2(u)| < \infty \quad .$$

The Bounded Convergence Theorem guarantees that both estimates (2.25) and (2.27) are asymptotically unbiased i.e.

$$E[f_n(x)] \to f(x) \ , \quad E[f_n^{(1)}(x)] \to f^{(1)}(x), \text{ as } n \to \infty.$$

We may now form the estimate

$$\alpha_n(x) = (v(x) - \theta_0) f_n(x) - f_n^{(1)}(x) \tag{2.29}$$

and $T = \{t_n\}$ is given by (2.6).

Lemma 2.3

The excess risk satisfies

$$0 \leqslant R_n(T,G) - R(G) \leqslant \int_a^\infty |\alpha(x)| \Pr\{|\alpha_n(x) - \alpha(x)| \geqslant |\alpha(x)|\} dx$$

$$\leqslant \int_a^\infty |\alpha(x)|^{1-\delta} |\alpha_n(x) - \alpha(x)|^\delta dx \quad . \qquad (2.30)$$

The proof of this is the continuous analogue of Theorem 2.1. The next result stated without proof gives general conditions for rates of convergence in the exponential family.

Theorem 2.4

Let $f_\theta(x)$ be defined by (2.20) with $t_n(x)$ and $\alpha_n(x)$ given by (2.6) and (2.29) respectively. The estimates $f_n(x)$, $f_n^{(1)}(x)$ are defined by (2.25) and (2.27). If $h^{(r)}(x)$ exists for some integer $r \geqslant 2$ (for $x > a$), and if for some δ, $0 < \delta < 2$, and some $\varepsilon > 0$

$$\int_a^\infty |\alpha(x)|^{1-\delta} (1+|v(x)|)^\delta \{f_\varepsilon^*(x)\}^{\delta/2} dx < \infty \qquad (2.31)$$

where $f_\varepsilon^*(x) = \sup_{0 \leqslant t \leqslant \varepsilon} \{f(x+t)\}$,

$$\int_a^\infty |\alpha(x)|^{1-\delta} (1 + v(x))^\delta \{q_\varepsilon^{(r)}(x)\}^\delta dx < \infty \quad , \qquad (2.32)$$

where $q_\varepsilon^{(r)}(x) = \sup_{0 \leqslant t \leqslant \varepsilon} |f^{(r)}(x+t)|$

then, choosing $a_n = 0\left(n^{-\frac{1}{2r+1}}\right)$ and kernels K_1 and K_2 satisfying (2.26) and (2.28) respectively such that

$$\int_0^{u_i} u^{i+j-1} K_i(u)\,du = 0$$

for $i = 1,2$, $j = 1,2, \ldots, r-1$, provides that $T = \{t_n\}$ is asymptotically optimal of order $n^{-\gamma}$, where $\gamma = (r-1)\delta/(2r+1)$.

This result depends on the following, derived from the c_δ - inequality (Loeve, p155).

$$E|\alpha_n(x) - \alpha(x)|^\delta \leqslant c_\delta^2(|\theta_0| + |v(x)|^\delta)E|f_n(x) - f(x)|^\delta + c_\delta E|f_n^{(1)}(x) - f^{(1)}(x)|^\delta \quad .$$

The remainder of the proof consists in finding rate bounds for $E|f_n(x) - f(x)|^\delta$ and $E|f_n^{(1)}(x) - f^{(1)}(x)|^\delta$.

This theorem is basically that of Johns and Van Ryzin (1972). Their additional assumptions that $\int_\Omega \theta dG(\theta) < \infty$ and that $h^{(r)}(x)$ is continuous, while not being unduly restrictive, are not necessary for the proof of the above result. The conditions (2.31) and (2.32) are unintuitive and the following example illustrates that these can be reduced to simple moment conditions on the a priori distribution.

Example

Consider the simple scale exponential density

$$f_\theta(x) = \begin{cases} \theta e^{-\theta x} & , \quad \theta > 0 \ , \ x > 0 \\ 0 & , \quad \text{elsewhere} \end{cases} \quad . \tag{2.33}$$

This density is a member of the family (2.20) with $h(x) = 1$ if $x > 0$ and $h(x) = 0$ if $x \leqslant 0$, and $\beta(\theta) = \theta$ for $\theta > 0$. Since $f_\theta(x)$ is monotonically decreasing in x, so too is $f(x)$ and thus

$$f_\varepsilon^*(x) = \sup_{0 \leqslant t \leqslant \varepsilon} \{f(x+t)\} = f(x)$$

for all x, for all $\varepsilon > 0$. Similarly

$$|f^{(r)}(x)| = \int_\Omega \theta^r f_\theta(x) dG(\theta) \quad \text{is decreasing}$$

to give

$$|q_\varepsilon^{(r)}(x)| = |f^{(r)}(x)| \ , \quad \forall\ x, \quad \forall\ \varepsilon > 0 \ .$$

This leads to the next result.

Corollary:

For the scale exponential (2.33), the hypotheses of Theorem 2.4 hold for each $0 < \delta \leqslant 1$ if

$$\int_\Omega \theta^r dG(\theta) < \infty \tag{2.34}$$

$$\int_\Omega \theta^{-\eta} dG(\theta) < \infty \ , \text{ with } \eta = (1+t)\delta/(2-\delta) \tag{2.35}$$

for some $t > 0$. The proof of this result may be found in Yu (1971) who gives a slight improvement of Johns and Van Ryzin's result.

When we assume that the prior is a member of a certain parametric family the verification of the conditions of the

theorem becomes somewhat easier. Suppose that in the above corollary we assume that $G(\theta)$ is a prior distribution having the density

$$G'(\theta) = g(\theta) = \begin{cases} \theta^t e^{-\alpha\theta} q(\theta) & , \quad \theta > 0, \alpha > 0, \ t > 0 \\ 0 & , \quad \text{otherwise} \end{cases}$$

where $q(\theta)$ is bounded, by B say, on $(0,\infty)$. Putting $\delta = 1$ and $p = t/2 > 0$ gives

$$\int_0^\infty \theta^{-(1+p)\delta/(2-\delta)} dG(\theta) < \infty$$

which gives (2.35). Also, for $r \geqslant 2$

$$\int_0^\infty \theta^r dG(\theta) = \int_0^\infty \theta^{t+r} e^{-\alpha\theta} q(\theta) d\theta$$

$$\leqslant B \frac{\Gamma(t+r+1)}{\alpha^{t+r+1}} < \infty \quad .$$

Thus we are guaranteed of a rate of convergence of $0\left(n^{-\frac{r-1}{2r-1}}\right)$.

Yu (1971) has shown that the $N(-\theta,1)$ density is a member of the family (2.20) and that the conditions of Theorem 2.4 are satisfied for $0 < \delta \leqslant 1$ provided that

$$\int_\Omega \theta^{1+(2+t)\delta/(2-\delta)} dG(\theta) < \infty$$

for some $t > 0$.

Remarks

Johns and Van Ryzin (1972) have also investigated the family

$$f_\theta(x) = \begin{cases} \theta^x \beta(\theta) h(x) & , \quad x > a \text{ , } \theta \text{ in a subinterval of } (0,\infty) \\ 0 & , \quad \text{otherwise} \end{cases}$$

where a may be finite or infinite and $h(x) > 0$ for $x > a$. Samuel (1963) has shown that some of the commonly encountered distributions are members of this family. A rate result very similar to Theorem 2.4 has been given in the Johns and Van Ryzin paper. Some improvements in these results have been made. Meeden (1972) points out that the empirical Bayes rules formulated by Johns and Van Ryzin are inadmissible. In fact, since the exponential family (2.20) has a monotone likelihood ratio and since the problem itself is monotone; i.e.

$$L(a_0,\theta) - L(a_1,\theta) \leqslant 0 \quad , \theta < \theta_0$$

$$L(a_0,\theta) - L(a_1,\theta) \geqslant 0 \quad , \theta \geqslant \theta_0 \ ,$$

then the class C, of all monotone decision procedures is essentially complete. [That is, given any procedure not in C, we can find one, in C, which is at least as good.] For a proof of this see Ferguson (1967). In this case, the class of monotone decision procedures is

$$C = \{t_c(x) = \begin{cases} a_1 \text{ if } x \geqslant c \\ \\ a_0 \text{ if } x < c \ . \end{cases} \} \ .$$

If we write

$$\Delta_G(x) = \int_{\Omega} [L(a_0,\theta) - L(a_1,\theta)] f_\theta(x) dG(\theta)$$

then the corresponding Bayes procedure is written as

$$t_G(x) = \begin{cases} 1 \ , & \text{if } \Delta_G(x) \leqslant 0 \\ 0 \ , & \text{if } \Delta_G(x) > 0 \end{cases} ,$$

[where, as before $t(x) = \Pr(\text{accept } H_0)$]. The empirical Bayes rule of Johns and Van Ryzin basically estimates $\Delta_G(x)$ by $\Delta_n(x)$ and replaces $t_G(x)$ by $t_n(x)$, where

$$t_n(x) = \begin{cases} 1 \ , & \text{if } \Delta_n(x) \leqslant 0 \\ 0 \ , & \text{if } \Delta_n(x) > 0 \end{cases} .$$

Van Houwelingen (1976) has given a method for constructing a test of the form

$$t_n^*(x) = \begin{cases} 1 \ , & \text{if } x < c_n(x_1, \ \ldots \ , x_n) \\ 0 \ , & \text{if } x \geqslant c_n(x_1, \ \ldots \ , x_n) \end{cases} ,$$

to give a weakly admissible rule in the sense that $t_n^*(x)$ is an admissible test for the non-empirical Bayes problem for all $x_1, \ \ldots \ , x_n$ for all n.

The basis of the convergence rate theorems in this section is the estimation of probability densities using kernel functions K. Examples of this include Rosenblatt (1956), Parzen (1962), Bhattacharya (1967) and Singh (1977). While these can provide excellent rate results, the restrictions imposed by Johns and Van Ryzin on the kernels are very severe. Consequently it is difficult to find functions K_1, K_2 which satisfy Theorem 2.4. An example of a function which does

meet the requirements is a polynomial of degree r-1, found upon the solution of the simultaneous equations formed by the conditions

$$\int_0^{u_i} u^{i+j-1} K_i(u)\,du = 0 \qquad i = 1,2;\ j = 1, \ldots, r-1.$$

However, as r increases the difficulty of solving such equations becomes considerably greater and as a consequence of this the convergence rate is not as good as the theory would permit.

Also, when considering the applicability of such results it is not sufficient to determine an asymptotic rate of convergence. When looking at $O(n^{-\alpha})$ convergence it is also necessary to look at the bounding constant in order to determine a result of the form (2.1). Unfortunately these kernels can give rise to very large bounds and as a result the number of observations needed to give a required degree of accuracy tend to be rather large. For a detailed analysis of this see Mara (1975).

Empirical Bayes methodology has also been applied to the multiple decision problem. i.e. $A = \{a_0, \ldots, a_k\}$. Deely (1965) gives asymptotically optimal decision procedures for several examples where a parametric form is assumed on the prior G and for the case where this restriction is removed. Van Ryzin (1970) and more recently Van Ryzin and Susarla (1977) have studied rates of convergence for this problem but as they are similar to the two-action case they will not be discussed further here.

(b) Estimation

Since the Bayes estimator $t_G(x)$ for the squared error loss problem is the posterior mean of θ given $X = x$, it is well-defined. Further by (2.23)

$$t_G(x) = v(x) - f^{(1)}(x)/f(x) \quad . \tag{2.36}$$

Hence estimation of $t_G(x)$ reduces to estimation of the unconditional density $f(x)$ and its derivative. Using the estimates (2.25) and (2.27) we can obtain an empirical Bayes estimator $t_n(x)$ similar in form to (2.36). As for the discrete case however, we must allow for the case $f_n(x) = 0$.

Let $\eta > 0$ and define

$f_n'(x) = \max\{f_n(x), \eta\}$ to produce the estimator

$$t_n'(x) = v(x) - f_n^{(1)}(x)/f_n'(x) \quad . \tag{2.37}$$

Theorem 2.5

Let $f_\theta(x)$ be of the form (2.20). Let $T = \{t_n'\}$ be a sequence of empirical Bayes procedures with $t_n'(x)$ defined by (2.37). If $h^{(r)}(x)$ exists for $x > a$ and if for some $\varepsilon > 0$

$$\int_a^\infty \{1 + [f^{(1)}(x)/f(x)]^2\} f_\varepsilon^*(x) f(x) dx < \infty$$

$$\int_a^\infty \{1 + [f^{(1)}(x)/f(x)]^2\} (q_\varepsilon^{(r)}(x))^2 f(x) dx < \infty$$

where $f_\varepsilon^*(x)$ and $q_\varepsilon^{(r)}(x)$ are given in Theorem 2.4, and if there exists $\delta > 0$ such that

$$\int_{\{x:\ f(x)\ <\eta\}} \{f^{(1)}(x)/f(x)\}^2 f(x)dx \leqslant c_1\eta^{\delta}$$

for some $c_1 > 0$, then choosing $a_n = 0\left(n^{-1/(2r-1)}\right)$ and $\eta = n^{-\gamma}$, with $\gamma = 2(r-1)/(2r+1)(2+\delta)$, the sequence T is asymptotically optimal with

$$0 \leqslant R_n(T,G) - R(G) = 0(n^{-\gamma'})$$

where $\gamma' = \gamma\delta = 2\delta(r-1)/(2r+1)(2+\delta)$.

Corollary 1

For the scale exponential density (2.33) the hypotheses of Theorem 2.5 hold for each $r \geqslant 2$ and $0 < \delta < 1$ provided only that

$$\int_{\Omega} \theta^{r+\frac{1}{2}} dG(\theta) < \infty \quad .$$

The above results are basically those of Yu (1971). These have been extended in Lin (1975) who considers the more general case of estimating a function of the parameter. He also gives a rate theorem for the scale exponential family when the prior distribution has the gamma density

$$g(\theta) = \theta^{\alpha-1}e^{-\theta}/\Gamma(\alpha) \qquad , \qquad \theta > 0 \quad ,$$

for some $0 < \alpha < \infty$. Unfortunately both the rate of convergence and the constant bound associated with it involve the value α, so that exact knowledge of the prior is again

required before we can determine any numerical results. The general comments of the previous section concerning the estimation of f(x) by kernel functions also apply here.

Since the theoretical results of this chapter give poor small sample properties it may be useful to consider other methods of finding empirical Bayes rules. Maritz (1966, 1967, 1970) has analysed selected examples using computer simulation of the experiments and the numerical results he gives for both the ordinary empirical Bayes procedures and his 'smooth' procedures do indicate that good small sample properties can be obtained. The next section considers empirical Bayes rules when the a priori distribution is a member of a given parametric family. The possible advantage of this is that the Bayes rule may admit a simpler estimator than the nonparametric ones of this section.

II. RATES FOR PARAMETRIC FAMILIES

The general rate of convergence results of the last section were derived without any parametric assumptions on the a priori distribution $G(\theta)$. However, as already noted, when determining their applicability it was necessary to specify an exact prior in order to get numerical results. In view of this, it would therefore seem reasonable to consider what sort of empirical Bayes rules and rates of convergence we can get if it is assumed from the beginning that G is a member of some parametric family. The prior distributions to be considered here will be the natural conjugates of the

density $f_\theta(x)$. The concept of conjugate distributions has been introduced and developed by Raifa and Schaifer (1961).

Since for both the estimation problem under squared error loss and for the two-action problem under the piecewise linear loss the Bayes rule $t_G(x)$ is a function of the posterior mean $E[\theta|X = x]$, we shall consider estimators for this function in the following examples.

Example 1. Poisson - Gamma

Suppose that the conditional distribution of X given θ is the Poisson distribution with parameter θ. Suppose further that the a priori distribution has a gamma density

$$g(\theta) = \frac{\beta^\alpha}{\gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \quad , \qquad \theta > 0,\ \alpha > 0,\ \beta > 0 \qquad (2.38)$$

where β will be assumed known. For this case the posterior distribution of θ given $X = x$ is also a gamma distribution with parameters $\alpha + x$ and $\beta + 1$.

The posterior mean in this case is therefore

$$E[\theta|X = x] = \frac{\alpha+x}{\beta+1} \quad . \qquad (2.39)$$

Consequently an estimate for α gives an estimator for the posterior mean, which for the estimation and two action problems mentioned before, will therefore give an empirical Bayes procedure. Note that since $E[X_i|\theta] = \int_{\mathcal{X}} x f_\theta(x)dx = \theta$ we have

$$E[X_i] = E[E[X_i|\theta]] = E[\theta] = \frac{\alpha}{\beta} \quad , \quad \forall i \quad .$$

Thus a simple estimator for α, using the previous observations $X_1, \ldots, X_n$ and the present observation $X = X_{n+1}$, is now

$$\hat{\alpha}_n = \beta \bar{X}_{n+1} \quad , \quad \text{where } \bar{X}_{n+1} = \frac{1}{n+1} \sum_{i=1}^{n+1} X_i \quad ,$$

and by the Law of Large numbers $\hat{\alpha}_n \xrightarrow{p} \alpha$.

If we now consider the squared-error loss estimation problem, the Bayes estimator is exactly the posterior mean and the corresponding empirical Bayes estimator is

$$t_n(x) = t_n(x_1, \ldots, x_n; x) = \frac{\hat{\alpha}_n + x}{\beta + 1} \quad . \tag{2.40}$$

We have $t_n(x) \xrightarrow{p} t_G(x)$ for all x and if $T = \{t_n\}$,

$$0 \leqslant R_n(T,G) - R(G) = E[t_n - t_G]^2$$

$$= E\left|\frac{1}{\beta+1} (\hat{\alpha}_n - \alpha)\right|^2$$

$$= \frac{\beta^2}{(\beta+1)^2} E\left|\frac{1}{n+1} \sum_{i=1}^{n+1} X_i - \frac{\alpha}{\beta}\right|^2$$

$$= \frac{\beta^2}{(\beta+1)^2} \cdot \frac{\alpha(\beta+1)}{(n+1)\beta^2} = \frac{\alpha}{(\beta+1)} \cdot \frac{1}{n+1} \quad . \tag{2.41}$$

This is a much better convergence rate than those given by the general estimators of the previous section even when they were restricted to the parametric case. Further, since the Bayes risk for this problem is $R(G) = \alpha/\beta(\beta+1)$, the relative error of the estimator (2.40) is

$$\frac{\alpha\beta(\beta+1)}{\alpha(\beta+1)\cdot n+1} = \frac{\beta}{n+1} .$$

Example 2. Normal - Normal

The conditional distribution of X given θ is $N(\theta,\beta^2)$. When the prior distribution is normal as well, say $N(\mu,\sigma^2)$, then it is well known that the posterior distribution of θ given $X = x$ is $N\left(\frac{\beta^2\mu + \sigma^2 x}{\beta^2+\sigma^2}, \frac{\sigma^2\beta^2}{\beta^2+\sigma^2}\right)$. We will again consider the squared error loss estimation problem.

Case 1. σ^2 known, μ unknown.

Since the unconditional distribution of the X_i is $N(\mu, \sigma^2 + \beta^2)$ we estimate μ by

$$\hat{\mu}_n = \frac{1}{n+1}\sum_{i=1}^{n+1} X_i .$$

Also, since

$$t_G(x) = \frac{\beta^2\mu + \sigma^2 x}{\sigma^2 + \beta^2} , \tag{2.42}$$

we form the empirical Bayes estimator

$$t_n(x) = \frac{\beta^2\hat{\mu}_n + \sigma^2 x}{\sigma^2 + \beta^2} . \tag{2.43}$$

This yields

$$0 \leq R_n(T,G) - R(G) = E[t_n - t_G]^2$$

$$= E\left|\frac{\beta^2(\hat{\mu}_n - \mu)}{\sigma^2+\beta^2}\right|^2$$

$$= \frac{\beta^4}{\sigma^2 + \beta^2} \cdot \frac{1}{n+1} \quad . \tag{2.44}$$

Furthermore since the Bayes risk is $\sigma^2\beta^2/(\sigma^2+\beta^2)$ we have a relative error of $\beta^2/\sigma^2(n+1)$.

Case 2. μ known, σ^2 unknown.

Rewriting (2.42) we get

$$t_G(x) = \beta^2 \cdot \frac{\mu}{\sigma^2+\beta^2} + \left[1 - \frac{\beta^2}{\sigma^2+\beta^2}\right]x \quad . \tag{2.45}$$

So we need an estimator for $1/(\sigma^2+\beta^2)$. Consider the unbiased estimator $(n-1)/s^2$, where $s^2 = \sum_{i=1}^{n+1} (X_i-\mu)^2$. This gives the empirical Bayes estimator

$$t_n(x) = \mu + (1 - \beta^2 \cdot \frac{n-1}{s^2})(x-\mu) \tag{2.46}$$

and thus

$$0 \leq R_n(T,G) - R(G) = E[\beta^4 \left(\frac{n-1}{s^2} - \frac{1}{\sigma^2+\beta^2}\right)^2 (X-\mu)^2]$$

$$= \frac{\beta^2}{n+1} E\left[\left(\frac{n-1}{s^2} - \frac{1}{\sigma^2+\beta^2}\right)^2 s^2\right] \tag{2.47}$$

[by independence of $X_1, \ldots, X_n, X_{n+1}$.] Note also that since $s^2/(\sigma^2+\beta^2)$ has a χ^2_{n+1} distribution (2.47) becomes

$$= \frac{\beta^4}{n+1} E\left[\frac{(n-1)^2}{s^2} - \frac{2(n-1)}{\sigma^2+\beta^2} + \frac{s^2}{(\sigma^2+\beta^2)^2}\right]$$

$$= \frac{2\beta^4}{(\sigma^2+\beta^2)(n+1)} \quad .$$

The relative error is now $2\beta^2/\sigma^2(n+1)$.

Case 3. μ unknown, σ^2 unknown.

Using the natural estimators

$$\hat{\mu}_n = \frac{1}{n+1}\sum_{i=1}^{n+1} X_i \quad , \quad \left(\frac{\hat{1}}{\sigma^2+\beta^2}\right)_n = \frac{n-2}{\sum_{i=1}^{n+1}(X_i-\hat{\mu}_n)^2}$$

we form the empirical Bayes estimator

$$t_n(x) = x - \beta^2\left(\frac{n-2}{s^2}\right)(x-\hat{\mu}_n) \quad , \tag{2.48}$$

where $s^2 = \sum_{i=1}^{n+1}(X_i - \hat{\mu}_n)^2$. Hence

$$0 \leqslant R_n(T,G) - R(G) = \beta^4 E\left[\frac{n-2}{s^2}(x-\hat{\mu}_n) - \frac{1}{\sigma^2+\beta^2}(x-\mu)\right]^2$$

$$= \beta^4 E\left[\left(\frac{n-2}{s^2} - \frac{1}{\alpha^2+\beta^2}\right)(x-\hat{\mu}_n) + \frac{1}{\sigma^2+\beta^2}(\mu-\hat{\mu}_n)\right]^2$$

and since $\hat{\mu}_n$ and s^2 are independent,

$$= \beta^4\{E\left[\frac{1}{n+1}\left(\frac{n-2}{s^2} - \frac{1}{\sigma^2+\beta^2}\right)^2 s^2\right] + \frac{1}{(\sigma^2+\beta^2)^2}E[\mu-\hat{\mu}_n]^2\}$$

$$= \beta^4\left\{\frac{2}{(\sigma^2+\beta^2)(n+1)} + \frac{1}{(\sigma^2+\beta^2)(n+1)}\right\}$$

$$= \frac{3\beta^4}{(\sigma^2+\beta^2)(n+1)} \quad .$$

We note here that the above results may be generalised to the case where r observations are taken in each X_i,

$X_i = (X_{i,1}, \ldots, X_{i,r})$. The Bayes procedure is

$$t_G(x) = \frac{\sigma^2\bar{x} + \beta^2\mu/r}{\sigma^2 + \beta^2/r} \quad , \quad \bar{x} = \frac{1}{r}\sum_{j=1}^{r} x_{n+1,j}$$

and the corresponding Bayes risk is $R(G) = \sigma^2\beta^2/(\sigma^2+r\beta^2)$.

Example 3. Exponential - Gamma

Suppose X has the scale exponential density with parameter θ

$$f_\theta(x) = \begin{cases} \theta e^{-\theta x} & , \quad \theta > 0 \ , \ x > 0 \\ 0 & , \quad \text{elsewhere} \ . \end{cases}$$

Assume that the prior distribution has the gamma density

$$g(\theta) = \begin{cases} \dfrac{\beta^\alpha \theta^{\alpha-1} e^{-\beta\theta}}{\Gamma(\alpha)} & , \quad \theta > 0 \\ 0 & , \quad \text{elsewhere} \ , \end{cases}$$

and further assume α is known, β is unknown. Since the posterior distribution of θ given $X = x$ is gamma with parameters $\alpha + 1$ and $\beta + x$ we have for the squared error loss estimation problem

$$t_G(x) = \frac{\alpha+1}{\beta+x} \quad .$$

Since

$$E[X_i] = E[E(X_i|\theta)] = E[\frac{1}{\theta}] = \frac{\beta}{\alpha-1} \quad , \quad \text{for } \alpha > 1 \ ,$$

we have a natural estimator for β as

$$\hat{\beta}_n = (\alpha-1)\frac{1}{n+1}\sum_{i=1}^{n+1} X_i$$

to give $t_n(x) = (\alpha+1)/(\hat{\beta}_n + x)$.

Hence the empirical Bayes error is

$$E[t_n - t_G]^2 = (\alpha+1)^2 E\left[\frac{\hat{\beta}_n-\beta}{(\beta+x)(\hat{\beta}_n+x)}\right]^2$$

$$\leqslant \frac{(\alpha+1)^2}{\beta^2} E(\hat{\beta}_n-\beta)^2 \cdot E\left[\frac{1}{\hat{\beta}_n^2}\right] \quad . \tag{2.49}$$

To evaluate $E[\hat{\beta}_n^{-2}]$, we first notice that since the conditional distribution of X is exponential with parameter θ, the distribution of $\hat{\beta}_n = \frac{\alpha-1}{n+1} \sum_{i=1}^{n+1} X_i$ is gamma with parameters $n+1$ and $(n+1)\theta/(\alpha-1)$. Thus

$$E\left[\frac{1}{\hat{\beta}_n^2}\middle|\theta\right] = \left[\frac{(n+1)\theta}{\alpha-1}\right]^2 /n(n-1), \text{ giving}$$

$$E\left[\frac{1}{\hat{\beta}_n^2}\right] = \frac{(n+1)^2}{(\alpha-1)^2 n(n-1)} E[\Theta^2] = \frac{(n+1)^2\alpha(1+\alpha)}{(\alpha-1)^2 n(n-1)\beta^2} \quad .$$

Further

$$\text{Var}(X) = \frac{\beta^2\alpha}{(\alpha-1)^2(\alpha-2)} \quad , \quad \text{for} \quad \alpha > 2 \quad ,$$

and since $E[\hat{\beta}_n - \beta]^2 = (\alpha-1)^2 \cdot \frac{1}{n+1} \text{Var}[X]$ we have from (2.49), provided $\alpha > 2$

$$0 \leqslant R_n(T,G) - R(G) \leqslant \frac{1}{n+1} \cdot \frac{(\alpha+1)^2}{\beta^2} \cdot \frac{(\alpha-1)^2\cdot\beta^2\alpha}{(\alpha-1)^2(\alpha-2)} \cdot \frac{(n+1)^2\alpha(1+\alpha)}{(\alpha-1)^2 n(n-1)\beta^2}$$

$$= \frac{1}{n+1} \cdot \left[\frac{\alpha^2(\alpha+1)^3(n+1)^2}{(\alpha-1)^2(\alpha-2)n(n-1)\beta^2}\right] \quad ,$$

which is an $O\left(\frac{1}{n}\right)$ convergence rate. This compares with the $O(n^{-q})$, with $q = \alpha(r-1)/(3\alpha+2)(r+1)$, rate obtained by Lin (1975) for this example. His rate approaches $O(n^{-\frac{1}{3}})$ for r and α sufficiently large.

The above examples while not intended to be exhaustive provide an opportunity to illustrate that when we assume that we know G is in a particular parametric family, then we can more readily determine simple estimators for the unknown parameters and consequently obtain good rates of convergence to optimality. The nonparametric results given earlier are much more general from an asymptotic point of view, but since determination of their small sample properties seems to require specification of a parametric family for the prior distribution, and the small sample properties must be the criterion for assessing an empirical Bayes rule, then perhaps the above examples indicate a more useful approach.

CHAPTER III

EMPIRICAL BAYES WITH A SHIFTING PRIOR DISTRIBUTION

1. INTRODUCTION

The general empirical Bayes approach as described in Chapter I is concerned with making a decision about an unknown parameter θ based on an observation X whose density function is $f_\theta(x)$ and under the assumptions that;

(i) θ is a random variable with a priori distribution G; and

(ii) G is unknown, but we have the information from previous independent occurrences of the identical component problem. Using this information we can, under certain conditions, obtain desirable asymptotic properties for the empirical Bayes decision procedure.

Often, however, the history of the decision problem under discussion is such that the component problems vary with time and it is therefore of interest to determine whether or not asymptotic optimality can be recovered in such cases. An example where this model may be appropriate is the problem of determining the fraction defective θ in a consignment of resistors. The process which manufactures the resistors may produce defectives according to the distribution $G(\theta)$ at one time, but at some later time, perhaps due to wear and deterioration of the machinery, the distribution of the defectives follows $G'(\theta)$. If there is no relation between

$G(\theta)$ at time n and $G'(\theta)$ at time n', then we have no information from previous problems concerning the present one. That is, we have only one observation upon which to base our decision. If, however, the change in the component problems is a structured one then we may recover prior information to assist with the present problem.

Some early work in modifying the assumptions of empirical Bayes theory has been performed by Tainiter (1966), who has weakened the assumption that the X_i's are independent to one of r-dependence. That is, if B_n and C_n are the σ-fields generated by $\{X_n, X_{n+1}, \ldots\}$ and $\{X_1, \ldots, X_n\}$ respectively, then B_{n+r+1} is independent of C_n. He gives an asymptotically optimal procedure for the two-action problem under zero-one loss.

It is important to note here that if the component problem is different at each stage then the criterion of optimality must be modified accordingly. That is, at stage n+1 the Bayes rule will be $t_G^{(n+1)}(x)$ say, with corresponding Bayes risk $R_{n+1}(G)$. Hence if we use a procedure $t_n(x)$ based on $X_1, \ldots, X_n$ giving risk $R^{(n+1)}(t_n,G)$, the desired optimality goals would be

$$\text{(i)} \quad t_n(x) - t_G^{(n)}(x) \xrightarrow{p} 0 \qquad \text{a.e.}[\mu] \ , \tag{3.1}$$

and asymptotic optimality is now

$$\text{(ii)} \quad R^{(n+1)}(t_n,G) - R_{n+1}(G) \to 0 \ , \quad \text{as } n \to \infty \ . \tag{3.2}$$

Susarla and O'Bryan (1975) have examined the case where $G(\theta)$ remains fixed with support in $(0,\infty)$, but the observations are drawn according to the density

$$f_\theta(x) = Ke^{-K(x-\theta)} \quad , \qquad x > \theta \quad ,$$

where, at stage i, K = K(i) with {K(i)} being a known sequence of constants in a bounded positive interval. For the two action problem under the piecewise linear loss structure (1.32), an asymptotically optimal procedure is given together with a rate of convergence of $0(n^{-\delta/4})$ for those priors for which

$$P(\alpha \leqslant \theta \leqslant \alpha') \leqslant k(\alpha'-\alpha) \quad , \qquad \text{for all } \alpha \leqslant \alpha'$$

and

$$E[\theta^{(1+t)/(1-\delta)}] < \infty \ , \text{ for some } t > 0,\ 0 < \delta < 1 \ .$$

O'Bryan and Susarla (1975) have considered the modified version of the squared-error loss estimation of known linear functions of the unknown mean of normal distribution functions with known variances. Specifically X_n is distributed according to the $N(a_n\theta + b_n, \sigma_n^2)$ density, where $\{a_n\}$, $\{b_n\}$, $\{\sigma^2\}$ are known. For this normal case the Bayes estimator at stage n+1, is

$$t_G^{(n+1)}(x) = \frac{x - b_{n+1}}{a_{n+1}} + \sigma_{n+1}^2 \frac{f_{G,n+1}^{(1)}(x)}{f_{G,n+1}(x)} \quad ,$$

where

$$f_{G,n+1}(x) = \int_\Omega \frac{1}{\sqrt{2\pi}\sigma_{n+1}} e^{-\frac{1}{2}\left\{\frac{x - a_{n+1}\theta - b_n}{\sigma_n}\right\}^2} dG(\theta)$$

and $f_{G,n+1}^{(1)}(x)$ is the first order derivative of $f_{G,n+1}(x)$. By placing bounds on $\{a_n\},\{b_n\},\{\sigma_n^2\}$ O'Bryan and Susarla are

able to work within the standard empirical Bayes framework and, developing estimators for $f_{G,n+1}(x)$ based on its characteristic function, they give a procedure $t_n(x)$, estimating $t_{G,}^{(n)}(x)$, which is asymptotically optimal, for distributions G with a compact support, with a rate of convergence of $0((n^{-1}\log n)^{\gamma})$, for $0 < 2\gamma < 1$.

O'Bryan (1976) and O'Bryan and Susarla (1976) have modified the empirical Bayesian assumptions by allowing the number of observations taken at each stage to vary. That is, $X_i = (X_{i,1}, \ldots, X_{i,m(i)})$ is a sample of size m(i) taken from $f_\theta(x)$ at stage i. For the case where $f_\theta(x)$ is in the discrete exponential family (2.2) and provided that $m(i) \leqslant M < \infty$, O'Bryan exhibits an asymptotically optimal estimator for prior distributions with support on $[0,\beta]$ for some $\beta > 0$. When $f_\theta(x)$ is normal with mean θ, O'Bryan and Susarla give estimators which are asymptotically optimal with a rate of $0(n^{-a/(2+a)M})$ for $0 < a < 1$, providing G has support on [0,1].

The exponential family

$$f_{\theta,m}(x) = \begin{cases} e^{-\theta x}[Z(\theta)]^m h_m(x) & , \quad x > a \\ 0 & , \quad \text{elsewhere} , \end{cases}$$

has also been discussed by O'Bryan and Susarla (1977).

As an example of a member of this family they quote the normal distribution with mean $-m\theta$ and variance $m\sigma^2$, σ^2 known. At stage i, an observation X_i is drawn from the density $f_{\theta,m(i)}(x)$, where as before Ω is restricted to a compact set and $\{m(i)\}$ is known with $m(i) \leqslant M < \infty$. An

asymptotically optimal rule is given for the problem of squared error loss estimation.

All of the results mentioned above are concerned with modifying conditions on the observational data, while retaining the assumption that $G(\theta)$ is fixed throughout. Suppose now that, as mentioned in the example with the resistors, the prior distribution is not constant, but that it changes in some known way.

In particular consider the sequence (θ_1,X_1), (θ_2,X_2), ... , (θ_n,X_n) where as in the usual empirical Bayes case the $\{X_i\}$ are drawn independently, and given $\theta_i = \theta$, X_i is distributed according to the density $f_\theta(x)$, but now suppose that for each i, θ_i is drawn from the a priori distribution G_i. (It is assumed that all G_i's have the same support). This gives the unconditional density function of the observations as,

$$f_{G_n}(x) = \int_\Omega f_\theta(x)dG_n(\theta) \quad . \tag{3.3}$$

At stage n+1, (we have observed X_1, ... , X_n on $f_{G_1}(x)$, ... , $f_{G_n}(x)$ respectively), the a priori distribution is $G_{n+1}(\theta)$ and upon observing $X_{n+1} = x$ we form the corresponding Bayes procedure $t_{G_{n+1}}(x)$. The Bayes envelope functional is now

$$R(G_{n+1}) = R(t_{G_{n+1}}, G_{n+1})$$

$$= \int_\Omega \int_{\mathcal{X}} L(t_{G_{n+1}}(x),\theta) f_\theta(x)d\mu(x)dG_{n+1}(\theta) \quad . \tag{3.4}$$

This is the desired optimality goal at stage n+1 and if $G_{n+1}(\theta)$

is known then the decision problem is again the Bayesian one. If $G_{n+1}(\theta)$ is unknown then the performance of any empirical Bayes rule $t_n(x)$ based on $X_1, \ldots, X_n, X_{n+1}$ will be assessed relative to this criterion. Accordingly we have the following.

Definition

The sequence of decision procedures $T = \{t_n\}$ is said to be asymptotically optimal relative to $\{G_n\}$ if

$$\lim_{n\to\infty} \{R_n(T,G_{n+1}) - R(G_{n+1})\} = 0 . \tag{3.5}$$

II. GENERAL OPTIMALITY RESULTS

We now look to the extension of Robbins' general results on asymptotic optimality for fixed G to the case where $G = G_n$ at stage n. If we define $L(\theta) = \sup_{a\in A} L(a,\theta)$ and assume that the sequence of priors is such that

$$\int_\Omega L(\theta)dG_n(\theta) < \infty \quad , \qquad \text{for all } n \ , \tag{3.6}$$

then we have:

Theorem 3.1

The sequence $T = \{t_n\}$ is asymptotically optimal if (3.6) holds and if a.e.$[\mu]x$,

$$p\lim_{n\to\infty} \{\phi_{G_{n+1}}(t_n(x),x) - \phi_{G_{n+1}}(t_{G_{n+1}}(x), x)\} = 0 \tag{3.7}$$

where $\phi_{G_{n+1}}(a,x)$ is defined by (1.4).

The proof of this result is basically that of Lemma 1.1 and is therefore omitted. Further, if we define $\Delta_{G_{n+1}}(a,x)$ analagous to (1.23) and approximate it by a sequence $\Delta_n(a,x) = \Delta_n(x_1, \ldots, x_n; a,x)$ for which

$$p \lim_{n\to\infty} \sup_{a\in A} |\Delta_n(a,x) - \Delta_{G_{n+1}}(a,x)| = 0 \quad \text{a.e.}[\mu]x$$

then a result similar to Theorem 1.2 gives asymptotic optimality for $t_n(x) = a^* \in A$ such that

$$\Delta_n(a^*,x) \leqslant \inf_{a\in A} \Delta_n(a,x) + \varepsilon_n$$

for any sequence $\{\varepsilon_n\}$ converging to zero. The following corollary replaces (3.7) with a more intuitive condition.

Corollary

$T = \{t_n\}$ is asymptotically optimal if (3.6) holds and $L(a,\theta)$ and $t_n(x)$ are such that for all $\theta \in \Omega$

$$p \lim_{n\to\infty} \{L(t_n(x),\theta) - L(t_{G_{n+1}}(x),\theta)\} = 0 \qquad \text{a.e.}[\mu]x \ . \quad (3.8)$$

Proof

Suppose $B = \sup_x \sup_n f_{G_n}(x)$. Then $B < \infty$. Furthermore, for any $\varepsilon > 0$, $\exists\delta$ such that

$$1-\varepsilon \leqslant P(|L(t_n(x),\theta) - L(t_{G_{n+1}}(x),\theta)| < \delta)$$

$$\leqslant P(|\phi_{G_{n+1}}(t_n(x),x) - \phi_{G_{n+1}}(t_{G_{n+1}}(x),x)| < \delta B) \quad \text{a.e.}[\mu]$$

(by using (1.4)). That is $\phi_{G_{n+1}}(t_n(x),x) - \phi_{G_{n+1}}(t_{G_{n+1}}(x),x) \to 0$ in probability a.e.[μ]x.

We further note that (3.8) is guaranteed by the continuity of $L(a,\theta)$ in the first argument provided

$$p \lim_{n\to\infty} \{t_n(x) - t_{G_{n+1}}(x)\} = 0 \qquad \text{a.e.}[\mu]x. \tag{3.9}$$

As noted earlier (3.6) is a restrictive condition and while it holds for finite action problems under suitable moment conditions on the priors $\{G_n\}$, it is not valid for any decision problem where the action space is an unbounded interval, and in particular the squared error loss estimation problem on $(-\infty,\infty)$. Deely and Zimmer (1976) have weakened the overall boundedness condition on the loss function, requiring only a bounding integrable sequence on the losses $L(t_n(x),\theta)$. Since it will be used later, their result is now adapted for the changing prior case.

Theorem 3.2

Let $\{t_n(x)\}$ be a sequence of decision procedures for the modified empirical Bayes problem and suppose that $L(a,\theta)$ and $t_n(x)$ are such that (3.8) holds. Further, suppose there exists a sequence of functions $h_n(x,\theta) = h_n(x_1, \ldots, x_n; x,\theta)$ such that

(a) $p \lim_{n\to\infty} h_n(x,\theta) = h(x,\theta)$, for each (x,θ)

(b) $|L(t_n(x),\theta) - L(t_{G_{n+1}}(x),\theta)| \leq h_n(x,\theta)$

for each (x,θ) and for all n, and

$$(c) \quad \lim_{n\to\infty} E[h_n(x,\theta)] = E[\lim_{n\to\infty} h_n(x,\theta)] < \infty$$

where the expectation E is with respect to the random variables $X_1, \ldots, X_n, X = X_{n+1}$ and $\theta = \theta_{n+1}$.

Proof

Define $q_n(x,\theta) = |L(t_n(x),\theta) - L(t_{G_{n+1}}(x),\theta)|$.

Using Fatou's lemma for $q_n(x,\theta) \geqslant 0$

$$0 = E[\lim_{n\to\infty} q_n(x,\theta)] \leqslant \liminf_{n\to\infty} E[q_n(x,\theta)].$$

Also with $h_n(x,\theta) - f_n(x,\theta) \geqslant 0$ we have

$$h(x,\theta) = E[h(x,\theta)]$$

$$= E[\lim_{n\to\infty}(h_n(x,\theta) - q_n(x,\theta))]$$

$$\leqslant \liminf_{n\to\infty} E[h_n(x,\theta) - q_n(x,\theta)] \quad ,$$

by Fatou's lemma,

$$\leqslant h(x,\theta) - \limsup_{n\to\infty} E[f_n(x,\theta)] \quad ,$$

by condition (c). Thus,

$$0 \leqslant \liminf_{n\to\infty} E[q_n(x,\theta) \leqslant \limsup_{n\to\infty} E[q_n(x,\theta)] \leqslant 0 \quad .$$

That is $\lim_{n\to\infty} E[q_n(x,\theta)] = 0$ giving asymptotic optimality in the sense of (3.5).

III. ASYMPTOTICALLY OPTIMAL PROCEDURES

(1) Introduction

Suppose that the a priori distribution at stage i is $G_i(\theta)$ where

$$E_{G_i}[\theta] = \lambda_i \tag{3.10}$$

and

$$G_i'(\theta) = g_i(\theta) = g^*(\theta-\lambda_i)$$

for some fixed g^*. Suppose further that the conditional distribution of X_i, $f(x|\theta)$ is also a location parameter density with

$$f(x|\theta) = f_0(x-\theta) \quad . \tag{3.11}$$

This gives, for the ith observation

$$E[X_i] = E[E[X_i|\theta_i]] = E[\mu_0 + \theta_i] = \mu_0 + \lambda_i \tag{3.12}$$

where μ_0 is the mean of a random variable with probability density $f_0(t)$. Without loss of generality we may take $\mu_0 = 0$. Furthermore, the unconditional density function of X_i is

$$f_{G_i}(x_i) = \int_\Omega f(x_i|\theta_i) g_i(\theta_i) d\theta_i$$

$$= \int_{\Omega} f_0(x_i-\theta_i)g^*(\theta_i-\lambda_i)d\theta_i \quad . \tag{3.13}$$

Let $u_i = \theta_i - \lambda_i$, $du_i = d\theta_i$ and (3.13) becomes

$$\int_{\Omega} f_0(x_i-\lambda_i-u_i)g^*(u_i)du_i$$

$$= \int_{\Omega} f_0(y_i-u_i)g^*(u_i)du_i \quad , \quad \text{where } y_i = x_i-\lambda_i \ ,$$

$$= \int_{\Omega} f(y_i|u_i)g^*(u_i)du_i$$

$$= f_{G^*}(y_i) \quad , \quad \text{for all } i \ . \tag{3.14}$$

Thus, the random variables $\{Y_i\}$ are mutually independent, (since the $\{X_i\}$ are drawn independently) and by (3.14), are also identically distributed with

$$E[Y_i] = E[X_i - \lambda_i] = \lambda_i-\lambda_i = 0$$

$$Var[Y_i] = Var[X_i] = \sigma^2 \quad , \quad \text{say} \ .$$

For example if $f(x|\theta) = N(\theta,\sigma^2-1)$, $g_i(\theta) = N(\lambda_i,1)$ then $f_{G_i}(x_i)$ is $N(\lambda_i,\sigma^2)$, giving a $N(0,\sigma^2)$ density for all Y_i's.

The change in the means $\{\lambda_i\}$ is assumed to be of the form

$$\lambda_i = az_i + b \tag{3.15}$$

where $\{z_i\}$ is a known sequence and the parameters a,b are both unknown. Consequently on observing the random variable X_i on

$f_{G_i}(x)$, we have

$$X_i = \lambda_i + Y_i = az_i + b + Y_i \tag{3.16}$$

where the $\{Y_i\}$ are independent and identically distributed random variables with zero mean. This gives the conditions for the usual two-variable linear regression model. Hence, based on the observations $X_1, \ldots, X_k$, the least squares estimators for a and b are

$$a_k = \frac{\frac{1}{k}\sum_{i=1}^{k}(x_i-\bar{x})(z_i-\bar{z})}{\frac{1}{k}\sum_{i=1}^{k}(z_i-\bar{z})^2} \tag{3.17}$$

and

$$b_k = \bar{x} - a_k\bar{z} \quad , \tag{3.18}$$

where $\bar{x} = \frac{1}{k}\sum_{i=1}^{k} x_k$, $\bar{z} = \frac{1}{k}\sum_{i=1}^{k} z_i$. Thus we may form the double sequence of estimators for $\{\lambda_i\}$ as

$$\lambda_{i,k} = a_k z_i + b_k \quad . \tag{3.19}$$

Lemma 3.3

The estimates (3.19) are unbiased for each i and furthermore

$$E[\lambda_{i,k} - \lambda_i]^2 \leqslant \frac{2\sigma^2}{k}\left\{1 + \frac{(z_i-\bar{z})^2}{\frac{1}{k}\sum_{i=1}^{k}(z_i-\bar{z})^2}\right\} \quad . \tag{3.20}$$

Proof

Unbiasedness is established by taking simple expectations in (3.17) and (3.18). Next we have

$$E[\lambda_{i,k}-\lambda_i]^2 = E[a_k z_i+b_k - (az_i+b)]^2$$

$$= E[z_i(a_k-a) + (\bar{X}-a_k\bar{z}) - (\bar{\lambda}-a\bar{z})]^2$$

$$= E[\bar{X}-\bar{\lambda} + (z_i-\bar{z})(a_k-a)]^2$$

$$\leqslant 2\{E(\bar{X}-\bar{\lambda})^2 + (z_i-\bar{z})^2 E(a_k-a)^2\}$$

where $\bar{\lambda} = \frac{1}{k}\sum_{i=1}^{k} \lambda_i$. Since $E(a_k-a)^2 = \sigma^2 / \sum_{i=1}^{k} (z_i-\bar{z})^2$ [see Malinvaud (1970, p87)] the result follows.

When a is known, the squared error is simply σ^2/k. Also note that provided $\{z_i\}$ is bounded and

$$\frac{1}{k}\sum_{i=1}^{k} (z_i-\bar{z})^2 \to s^2 < \infty \qquad (3.21)$$

then the estimates $\lambda_{i,k}$ have a square mean convergence rate of $0(k^{-1})$ for each i. (Malinvaud, p87). This condition is assumed for the following work.

Suppose now we have observed $X_1, \ldots, X_n$. The 'present' prior is $G_{n+1}(\theta)$ and we observe $X = X_{n+1}$. The Bayes procedure is $t_{G_{n+1}}(x)$ and the corresponding risk functional is $R(G_{n+1})$. As noted in Chapter I, there are basically three methods of constructing empirical Bayes procedures and the method considered here is direct estimation of $t_{G_{n+1}}(x)$. In particular

it will be assumed that $t_{G_{n+1}}(x)$ can be characterised by the unconditional density $f_{G_{n+1}}(x)$ (and perhaps its first derivative).

We have noted that continuity of $L(a,\theta)$ together with

$$t_n(x) - t_{G_{n+1}}(x) \xrightarrow{p} 0 \qquad \text{a.e.}[\mu] \tag{3.22}$$

guarantees (3.8) which is a basic condition for the corollary to Theorem 3.1 and Theorem 3.2. Consequently if we can find estimators $f_n(x)$ and $f_n^{(1)}(x)$ for which

$$f_n^{(j)}(x) - f_{G_{n+1}}^{(j)}(x) \xrightarrow{p} 0 \qquad \text{a.e.}[\mu] \tag{3.23}$$

for $j = 0,1$, then (3.22) will be assured.

If the means $\{\lambda_i\}$ were known, then we could obtain the realisations of the independent and identically distributed random variables $Y_1 = X_1 - \lambda_1, \ldots, Y_n = X_n - \lambda_n$, $Y_{n+1} = Y = X_{n+1} - \lambda_{n+1}$, which places the problem in the usual empirical Bayes situation. Thus, from (3.14), we could construct estimators for the 'present' density $f_{G^*}(y)$ and its derivative $f_{G^*}^{(1)}(y)$, from which asymptotically optimal procedures may be constructed. However since the $\{\lambda_i\}$ are unknown we use the estimates $\lambda_{i,n}$ of (3.19) to form the random variables

$$Y_{i,n} = X_i - \lambda_{i,n}, \qquad i = 1, \ldots, n+1 \quad . \tag{3.24}$$

Therefore, at stage n+1, we observe $X_{n+1} = x$ and if we have

the consistent estimators $f_n(y)$, $f_n^{(1)}(y)$ for the 'present' density $f_{G^*}(y) = f_{G_{n+1}}(x)$ and its derivative $f_{G^*}^{(1)}(y) = f_{G_{n+1}}^{(1)}(x)$ respectively, based on $Y_1, \ldots, Y_n$ we replace Y_i by $Y_{i,n}$, for $i=1, \ldots, n$ and

$$\begin{aligned} Y_{n+1,n} &= X_{n+1} - \lambda_{n+1,n} \\ &= x - \lambda_{n+1,n} = y + (\lambda_{n+1} - \lambda_{n+1,n}) \\ &= \hat{y}, \text{ say.} \end{aligned} \tag{3.25}$$

This will yield the estimators

$$\begin{aligned} \hat{f}_n(x) &= f_{n,n}(y) \quad , \\ \hat{f}_n^{(1)}(x) &= f_{n,n}^{(1)}(y) . \end{aligned} \tag{3.26}$$

It must now be shown that such estimators are themselves consistent. That is

$$\begin{aligned} f_{n,n}(y) &\xrightarrow{p} f_{G^*}(y) \quad , \\ f_{n,n}^{(1)}(y) &\xrightarrow{p} f_{G^*}(y) \quad . \end{aligned} \qquad \text{a.e.}[y] . \tag{3.27}$$

(2) Empirical Density Estimators

Theorem 3.4

Let θ be a location parameter of the density $f(x|\theta)$ satisfying (3.11). Suppose we observe the continuous valued $X_1, \ldots, X_n, X_{n+1} = X$, each from $f(x|\theta_i)$, $i = 1, \ldots, n+1$ respectively, where θ_i has the a priori

distribution $G_i(\theta)$ satisfying (3.10) and (3.15). Let $Y_i = X_i - \lambda_i$, $i = 1, \ldots, n+1$, and let $Y_{i,n}$ be defined by (3.24).

With $\hat{y}$ given by (3.25), define

$$F_{n,n}(y) = \frac{\#\text{ of } Y_{1,n}, \ldots, Y_{n,n} \leqslant \hat{y}}{n} \quad . \tag{3.28}$$

Then,

$$F_{n,n}(y) - F_{G^*}(y) \xrightarrow{p} 0 \qquad \text{a.e. } [y] \quad . \tag{3.29}$$

Proof

Since $\lambda_{i,n} \xrightarrow{p} \lambda_i$, for each i as $n \to \infty$ this gives

$$Y_{i,n} = X_i - \lambda_{i,n} \xrightarrow{p} X_i - \lambda_i = Y_i \quad .$$

Define, for any $\varepsilon > 0$,

$$A_{n,\varepsilon} = \{|Y_{i,n} - Y_i| < \varepsilon \text{ for all } i=1, \ldots, n+1\} \quad .$$

On this set

$$F_{n,n}(y) \leqslant \frac{\#\text{ of } Y_1-\varepsilon, \ldots, Y_n-\varepsilon \leqslant \hat{y}}{n}$$

$$\leqslant \frac{\#\text{ of } Y_1, \ldots, Y_n \leqslant y + 2\varepsilon}{n} \quad ,$$

and

$$F_{n,n}(y) \geqslant \frac{\#\text{ of } Y_1+\varepsilon, \ldots, Y_n+\varepsilon \leqslant \hat{y}}{n}$$

$$\geqslant \frac{\#\text{ of } Y_1, \ldots, Y_n \leqslant y-2\varepsilon}{n} \quad .$$

That is,

$$F_n(y-2\varepsilon) \leqslant F_{n,n}(y) \leqslant F_n(y+2\varepsilon) \quad .$$

Since $F_n(y) \xrightarrow{a.e.} F_{G^*}(y)$ uniformly by the Glivenko-Cantelli Theorem and with

$$\lim_{n\to\infty} P(A_{n,\varepsilon}) = 1$$

we have, for any $\varepsilon > 0$,

$$F_{G^*}(y-2\varepsilon) - F_{G^*}(y) \leqslant p \lim_{n\to\infty} F_{n,n}(y) - F_{G^*}(y)$$

$$\leqslant F_{G^*}(y+2\varepsilon) - F_{G^*}(y) \quad .$$

Since F_{G^*} is continuous we have the result.

Recalling the estimates of Robbins (1963) for the continuous case

$$f_n(y) = \frac{F_n(y+c_n) - F_n(y-c_n)}{2cn}, \quad c_n \downarrow 0, \ nc_n \to \infty \ ,$$

$$f_n^{(1)}(y) = \frac{f_n(y+c_n) - f_n(y-c_n)}{2c_n} \ , \quad c_n \downarrow 0, \ nc_n^3 \to \infty \ ,$$

$$= \frac{F_n(y+2c_n) - F_n(y-2c_n)}{4c_n^2} \ ,$$

we form the corresponding estimates based on $Y_{1,n}, \ \ldots \ , Y_{n,n}, Y_{n+1,n} = \hat{y}$.

$$f_{n,n}(y) = \frac{F_{n,n}(y+c_n) - F_{n,n}(y-c_n)}{2c_n} \tag{3.30}$$

$$f^{(1)}_{n,n}(y) = \frac{F_{n,n}(y+2c_n) - F_{n,n}(y-2c_n)}{4c_n^2} \quad . \tag{3.31}$$

Robbins (1963) shows $f_n(y) \xrightarrow{p} f_{G^*}(y)$, $f^{(1)}_n(y) \xrightarrow{p} f_{G^*}(y)$, provided the derivative exists. From Theorem 3.5 we have:

Corollary

$$f_{n,n}(y) \xrightarrow{p} f_{G^*}(y) \qquad \text{a.e.[y]}. \tag{3.32}$$

$$f^{(1)}_{n,n}(y) \xrightarrow{p} f^{(1)}_{G^*}(y) \qquad \text{a.e.[y]}. \tag{3.33}$$

Proof

With $A_{n,\varepsilon}$ defined in Theorem 3.5 we have on this set

$$\frac{F_n(y+c_n-2\varepsilon) - F_n(y+2\varepsilon-c_n)}{2c_n} \leqslant \frac{F_{n,n}(y+c_n) - F_{n,n}(y-c_n)}{2c_n}$$

$$\leqslant \frac{F_n(y+c_n+2\varepsilon) - F_n(y-c_n-2\varepsilon)}{2c_n}$$

Since, by the Glivenko-Cantelli Theorem, the convergence of $F_n(y)$ to $F_{G^*}(y)$ is uniform in y and since F_{G^*} is continuous we get

$$f_{G^*}(y) \leqslant p \lim_{n\to\infty} \frac{F_{n,n}(y+c_n) - F_{n,n}(y-c_n)}{2c_n} \leqslant f_{G^*}(y) \quad .$$

That is (3.32) is valid. A similar proof is evident for (3.33).

(3) Kernel Function Estimators

The density estimators considered here are basically those of Johns and Van Ryzin (1972), since these have received considerable attention in empirical Bayes literature and may afford rate of convergence results. If we knew the values of the i.i.d. random variables $Y_1, \ldots, Y_n$, $Y_{n+1} = Y$ we could use as consistent estimators for $f_{G*}(y)$ and $f^{(1)}_{G*}(y)$ the functions

$$f_n(y) = \frac{1}{na_n} \sum_{j=1}^{n} K_1\left(\frac{Y_j - y}{a_n}\right) \tag{3.34}$$

and

$$f^{(1)}_n(y) = \frac{1}{na_n} \sum_{j=1}^{n} \left[\frac{1}{2a_n} K_2\left(\frac{Y_j - y}{2a_n}\right) - \frac{1}{a_n} K_2\left(\frac{Y_j - y}{a_n}\right)\right] \tag{3.35}$$

where $\{a_n\}$, $K_1(u)$, $K_2(u)$ satisfy (2.26) and (2.28). However, as $\{Y_j\}_1^{n+1}$ is unknown we replace Y_j by $Y_{j,n}$ and form the corresponding estimators $f_{n,n}(y)$ and $f^{(1)}_{n,n}(y)$. The following theorem imposes slightly stronger conditions on $\{a_n\}$ and the kernels K_1 and K_2 to guarantee consistency for these new estimates.

Theorem 3.5

Suppose θ is a location parameter of $f(x|\theta)$. We observe $X_1, \ldots, X_n, X_{n+1}(=X)$ on $f(x|\theta_1), \ldots, f(x|\theta_{n+1})$ respectively where θ_i has the a priori distribution $G_i(\theta)$ satisfying (3.10) and (3.15). Define $Y_i = X_i - \lambda_i$ and $Y_{i,n} = X_i - \lambda_{i,n}$ for $i=1, \ldots, n+1$, with $\lambda_{i,n}$ given by (3.19). Let $K_1(u)$ be a continuously differentiable function satisfying (2.26) where $\{a_n\}$ satisfies

$$a_n \downarrow 0, \quad na_n^4 \to \infty , \quad \text{as } n \to \infty .$$

Then, defining

$$f_{n,n}(y) = \frac{1}{na_n} \sum_{j=1}^{n} K_1\left(\frac{Y_{j,n}-\hat{y}}{a_n}\right) , \tag{3.36}$$

we have

$$f_{n,n}(y) \xrightarrow{p} f_{G^*}(y) \qquad \text{a.e.[y].} \tag{3.37}$$

Proof

Since K_1 is continuously differentiable on $[0,u_1]$, $\sup_{0\leqslant u\leqslant u_1} |K_1'(u)| = M_1$, say $M_1 < \infty$. That is $K_1 \in \text{Lip}_{M_1} 1$. In particular

$$\left| K_1\left(\frac{Y_j-y}{a_n}\right) - K_1\left(\frac{Y_{j,n}-\hat{y}}{a_n}\right)\right| \leqslant M_1 \left|\frac{Y_j-y-Y_{j,n}+\hat{y}}{a_n}\right|$$

for all $j = 1, \ldots, n$. Therefore with $f_n(y)$ defined by (3.34) we have

$$|f_n(y) - f_{n,n}(y)| \leqslant 2M_1 \sup_{1\leqslant j\leqslant n+1} \left|\frac{Y_j-Y_{j,n}}{a_n^2}\right| . \tag{3.38}$$

For any $\delta>0$, consider the set

$$A_N = \left\{ \left|\frac{Y_j-Y_{j,n}}{a_n^2}\right| < \delta/N, \quad n \geqslant N, \ j = 1, \ldots, n+1 \right\} .$$

On this set

$$|f_n(y) - f_{n,n}(y)| < \frac{2M_1\delta}{N} .$$

Further

$$1 \geqslant P(A_N) = P\left(1 \leqslant \sup_{j \leqslant n+1} \left| \frac{Y_j - Y_{j,n}}{a_n^2} \right| < \delta/N \right)$$

$$\geqslant 1 - E \left| \frac{Y_j - Y_{j,n}}{a_n^2} \right|^2 \Big/ (\delta/N)^2$$

for all j.

Since $E\,|Y_j - Y_{j,n}|^2 = E|\lambda_j - \lambda_{j,n}|^2 = O(\frac{1}{n})$ by the conditions on $\{z_j\}$, this gives

$$P(A_N) = 1 - O((Na_N^4)^{-1}).$$

That is $\lim_{N\to\infty} P(A_N) = 1$. Hence

$$f_{n,n}(y) - f_n(y) \xrightarrow{p} 0$$

and since $f_n(y) \xrightarrow{p} f_{G^*}(y)$, by the results of Johns and Van Ryzin (1972), we obtain the required result.

Consider now the estimator (3.35) for the first derivative of the density $f_G^*(y)$.

Theorem 3.6

Let the conditions of Theorem 3.6 be valid. Suppose $K_2(u)$ is a continuously differentiable function satisfying (2.28) with $\{a_n\}$ such that

$$a_n \downarrow 0, \quad na_n^6 \to \infty, \quad \text{as } n \to \infty \quad .$$

Define

$$f^{(1)}_{n,n}(y) = \frac{1}{na_n} \sum_{i=1}^{n} \left[\frac{1}{2a_n} K_2\left(\frac{Y_{j,n} - \hat{y}}{2a_n} \right) - \frac{1}{a_n} K_2\left(\frac{Y_{j,n} - \hat{y}}{a_n} \right) \right]. \qquad (3.39)$$

Then,

$$f^{(1)}_{n,n}(y) \xrightarrow{p} f^{(1)}_{G^*}(y) \qquad \text{a.e.[y]}. \qquad (3.40)$$

The proof of this result parallels that of Theorem 3.5 and is therefore omitted.

(4) The One-Parameter Exponential Family

To illustrate the results above we consider the family of densities (2.20), since for both the two-action linear loss problem and the squared-error loss estimation problem, the Bayes procedure may be expressed in terms of the density $f^{(x)}_{G_{n+1}}$, and its first derivative.

(a) Hypothesis Testing

We consider the test

$$H_o: \theta \leqslant \theta_o \text{ vs } H_1: \theta > \theta_o , \qquad (3.41)$$

with loss function (1.32). At stage n+1, the a priori distribution is $G_{n+1}(\theta)$ and the corresponding Bayes procedure is $t_{G_{n+1}}(x) = P(\text{accept } H_o | X=x)$, with

$$t_{G_{n+1}}(x) = \begin{cases} 1 & , \text{ if } \alpha_{G_{n+1}}(x) \leqslant 0 \\ 0 & , \text{ elsewhere} \end{cases}$$

where

$$\alpha_{G_{n+1}}(x) = (v(x) - \theta_o) f_{G_{n+1}}(x) - f^{(1)}_{G_{n+1}}(x) \quad .$$

Since for this problem $y = x - \lambda_{n+1}$, we define

$$v^*(y) = \frac{h^{(1)}(y+\lambda_{n+1})}{h(y+\lambda_{n+1})} = \frac{h^{(1)}(x)}{h(x)} = v(x) \quad . \tag{3.42}$$

Using (3.14) we have now

$$\alpha(y) = \alpha_{G_{n+1}}(x) = (v^*(y)-\theta_o) f_{G^*}(y) - f^{(1)}_{G^*}(y)$$

and the Bayes procedure at stage n+1 is now

$$t_{G_{n+1}}(x) = t_{G^*}(y) = \begin{cases} 1 \ , & \text{if } \alpha(y) \leqslant 0 \\ \\ 0 \ , & \text{elsewhere} \quad . \end{cases} \tag{3.43}$$

Thus if we had the values for $Y_1, \ldots, Y_n, Y_{n+1}$ we could use the estimates (3.30) and (3.31) or (3.34) and (3.35) to estimate $\alpha(y)$. However, using $Y_{1,n}, \ldots, Y_{n,n}, Y_{n+1,n}$ we have:

Theorem 3.7

Let θ be a location parameter of the density $f(x|\theta)$ in the family (2.20). Assume $h^{(1)}(x)$ exists and is continuous. With $f_{n,n}(y)$ and $f^{(1)}_{n,n}(y)$ given by (3.30) and (3.31) or (3.36) and (3.39) respectively, define $\alpha_n(y) = \alpha_n(y_{1,n}, \ldots, y_{n,n}; y) = \hat{\alpha}_n(x_1, \ldots, x_n; x)$ by

$$\alpha_n(y) = (v^*(y) - \theta_o) f_{n,n}(y) - f^{(1)}_{n,n}(y) \quad . \tag{3.44}$$

For the hypothesis testing problem (3.41) with loss (1.32) define

$$\hat{t}_n(x) = t_n(y) = \begin{cases} 1 & , \quad \text{if } \alpha_n(y) \leqslant 0 \\ \\ 0 & , \quad \text{if } \alpha_n(y) > 0. \end{cases} \tag{3.45}$$

Then $\hat{T} = \{\hat{t}_n\}$ is asymptotically optimal in the sense of (3.5) for all prior sequences for which

$$\int_\Omega \theta dG_n(\theta) < \infty \quad , \quad \text{for all } n.$$

Proof

Since $v^*(y)=v(x)$ is known, Theorems 3.5 - 3.7 guarantee that

$$\alpha_n(y) \xrightarrow{p} \alpha(y)$$

and therefore

$$t_n(y) \xrightarrow{p} t_{G^*}(y) \quad , \qquad \text{a.e.}[y].$$

With the assumption of a finite first order moment on the priors $\{G_n\}$, the conditions of Corollary 2 to Theorem 1.2 are satisfied ($\alpha(y) = -\Delta_{G^*}(y)$). Thus $T = \{t_n\}$ is asymptotically optimal relative to G^*. That is, $\hat{T} = \{\hat{t}_n\}$ is asymptotically optimal in the sense of (3.5).

(b) Estimation

It is well known that the Bayes estimator for the squared-error loss problem is the posterior mean. For the family (2.20) this is, at stage n+1,

$$t_{G_{n+1}}(x) = \frac{\int_\Omega \theta f(x|\theta) dG_{n+1}(\theta)}{f_{G_{n+1}}(x)}$$

$$= \frac{h^{(1)}(x)}{h(x)} - \frac{f^{(1)}_{G_{n+1}}(x)}{f_{G_{n+1}}(x)} \quad . \tag{3.46}$$

This may be re-expressed in the $y = x - \lambda_{n+1}$ variable as

$$t_{G^*}(y) = v^*(y) - \frac{f^{(1)}_{G^*}(y)}{f_{G^*}(y)} \quad . \tag{3.47}$$

As before, if we knew $Y_1, \ldots, Y_n, Y_{n+1}$, we could form consistent estimators $f_n(y)$ and $f^{(1)}_n(y)$ to give a consistent empirical Bayes rule. If we again assume that $h^{(1)}(x)$ exists and is continuous, then using (3.45) we form $t_n(y) = t_n(y_{1,n} \cdots, y_{n,n}; y) = \hat{t}_n(x_1, \ldots, x_n; x)$ as

$$t_n(y) = v^*(y) - \frac{f^{(1)}_{n,n}(y)}{f_{n,n}(y)} \quad . \tag{3.48}$$

with $f_{n,n}$ and $f^{(1)}_{n,n}$ given by (3.30) and (3.31) or (3.36) and (3.39). We now have a consistent empirical Bayes procedure.

However, as already noted, unboundedness of the action space precludes the use of Theorem 3.1. Following Deely and Zimmer (1976) we have the following:

Theorem 3.8

Let θ be a location parameter of $f(x|\theta)$, a unimodal and symmetric member of the family (2.20). Assume $h^{(1)}(x)$ is continuous. Suppose also that the a priori distributions are unimodal and symmetric and satisfy (3.10). Let $\hat{t}_n(x)$ be defined by (3.48). Let $s_n(x_1, \ldots, x_n; x)$ be

$$s_n(x) = \begin{cases} |x| + |\lambda_{n+1,n}| , & \text{if } \hat{t}_n(x) > |x| + |\lambda_{n+1,n}| \\ \hat{t}_n(x) , & \text{elsewhere} \\ -(|x| + |\lambda_{n+1,n}|), & \text{if } \hat{t}_n(x) < |x| + |\lambda_{n+1,n}| . \end{cases} \tag{3.49}$$

Then $T = \{s_n\}$ is asymptotically optimal in the sense of (3.5) for the problem of squared error loss estimation of θ.

Proof

Since $h^{(1)}(x)$ is continuous we have by (3.48) that $\hat{t}_n(x) - t_{G_{n+1}}(x) \xrightarrow{p} 0$ a.e., as $n \to \infty$. Also with $f(x|\theta)$ and $G_{n+1}(\theta)$ both unimodal and symmetric

$$|t_{G_{n+1}}(x)| = |E(\theta|x)| \leq |x| + |\lambda_{n+1}| .$$

Hence $s_n(x) - t_{G_{n+1}}(x) \xrightarrow{p} 0$, as $n \to \infty$, is assured from the fact that $\lambda_{n+1,n} - \lambda_{n+1} \xrightarrow{p} 0$. Now

$$|L(s_n,\theta) - L(t_{G_{n+1}},\theta)| \leq (|x| + |\lambda_{n+1,n}|)^2 + (|x| + |\lambda_{n+1}|)^2 + 2|\theta|(2|x| + |\lambda_{n+1,n}| + |\lambda_{n+1}|) . \tag{3.50}$$

Since $\{z_n\}$ is assumed to be a bounded sequence let

$$M = \sup_n \{\lambda_{n+1}, \lambda_{n+1,n}\} < \infty \quad .$$

The r.h.s. of (3.50) is thus bounded by

$$h(x,\theta) = 2(|x|+M)^2 + 4|\theta|(|x| + M) \quad .$$

The integrability of $h(x,\theta)$ is now guaranteed by the existence of second order moments for θ. Thus by Theorem 3.2 with $h_n(x,\theta) = h(x,\theta)$ for all n, for all (x,θ), $T = \{s_n\}$ is asymptotically optimal.

(5) Direct Estimation of the Prior

Thus far we have established asymptotic optimality by considering convergence in probability of the estimates a_n to a (and hence $\lambda_{i,n}$ to λ_i for each i). This convergence can be strengthened to a.e. convergence by the following. Since $x_i - \bar{x} = a(z_i - \bar{z}) + (y_i - \bar{y})$, where the $\{y_i\}$ are i.i.d. random variables with mean zero and finite variance σ^2, say, we rewrite (3.17) as

$$a_k = a + \frac{\sum_{i=1}^{k} (y_i - \bar{y})(z_i - \bar{z})}{\sum_{i=1}^{k} (z_i - \bar{z})^2}$$

$$= a + \frac{\sum_{i=1}^{k} y_i (z_i - \bar{z})}{\sum_{i=1}^{k} (z_i - \bar{z})^2}$$

By Theorem 4.3.1 of Lukacs (1975),

$$\frac{\sum_{i=1}^{k} Y_i(z_i-\bar{z})}{\sum_{i=1}^{k} (z_i-\bar{z})^2} \to 0 \qquad \text{a.e.}$$

provided

$$\lim_{k\to\infty} \sum_{i=1}^{k} (z_i-\bar{z})^2 = 0 \quad . \tag{3.52}$$

That is, $a_k \to a$ a.e. In this section it is assumed that (3.52) holds.

Lemma 3.9

Define $F_{n,n}(y)$ by (3.28) with $Y_{j,n}$ given by (3.19) and (3.24). Then

$$F_{n,n}(y) \xrightarrow{a.e.} F_{G^*}(y) \qquad \text{a.e.}[y] \quad . \tag{3.53}$$

Proof

Since $a_n - a \xrightarrow{a.e.} 0$, we have

$$|Y_{j,n} - Y_j| = |\lambda_{j,n} - \lambda_j| \xrightarrow{a.e.} 0 \ , \quad \text{for all } j.$$

Hence, for any $\varepsilon > 0$

$$P(\lim_{n\to\infty} |Y_{j,n} - Y_j| < \varepsilon, \text{ for all } j = 1, \ldots, n+1) = 1.$$

The rest of the proof parallels that of Theorem 3.4 and by use of the Glivenko-Cantelli Theorem we have the result.

When the Bayes procedure cannot be expressed in terms of the unconditional density $f_{G_{n+1}}(x)$ then it may be possible to construct a sequence $G_{n+1,n}(\theta)$ of estimators of $G_{n+1}(\theta)$ which converge weakly with probaility one. That is

$$P(\lim_{n\to\infty}(G_{n+1,n}(\theta) - G_{n+1}(\theta)) = 0 \text{ for all continuity points } \theta \text{ of } G_{n+1}) = 1 \qquad (3.54)$$

Since $G_{n+1}(\theta) = G^*(\theta-\lambda_{n+1}) = G(\theta')$, we need only find estimators of G^* which converge weakly with probability one.

Robbins (1964) has given a general method for determining such a sequence of estimators of G^* based on the empirical c.d.f. of the i.i.d. observations $Y_1, \ldots, Y_n$. Since these values are not available to us, we again use the random variables $Y_{1,n}, \ldots, Y_{n,n}$. Thus, given $F_{n,n}(y)$ as in (3.28), let $\mathcal{G}$ be a class of distributions in θ' which contains G^*. Let

$$d_n = \inf_{G\in\mathcal{G}} \sup_y |F_{n,n}(y) - F_G(y)| \quad .$$

Suppose $\hat{G}_n^* \in \mathcal{G}$, $\hat{G}_n^*(\theta') = \hat{G}_n^*(Y_{1,n}, \ldots, Y_{n,n}; \theta')$

is an element of $\mathcal{G}$ for which

$$\sup_y |F_{n,n}(y) - F_{\hat{G}_n^*}(y)| < d_n + \varepsilon_n \quad , \qquad (3.55)$$

where $\{\varepsilon_n\}$ is any sequence of constants tending to zero. Noting that $f(x|\theta) = f(y|\theta')$ gives $F(x|\theta) = F(y + \lambda_{n+1}|\theta')$ we have:

Theorem 3.10

Assume (i) $F(y|\theta')$ is continuous in θ'.

(ii) $\lim_{\theta' \to \infty} F(y|\theta')$ and $\lim_{\theta' \to -\infty} F(y|\theta')$ exist and are not distribution functions.

(iii) If $F_{G_1} = F_{G_2}$ then $G_1 = G_2$.

Then if $\{\hat{G}_n^*\}$ is defined by (3.55) we have

$$P(\lim_{n\to\infty} \hat{G}_n^*(\theta') = G^*(\theta'), \text{ for all continuity points } \theta' \text{ of } G^*) = 1 . \tag{3.56}$$

Using Lemma 3.9, the proof of this result is basically that of Robbins (1964) with the empirical c.d.f. $F_n(y)$ replaced by $F_{n,n}(y)$.

We note that condition (iii) is satisfied for the problem we are dealing with since $F(x|\theta)$ is a location parameter family. The estimators $\{\hat{G}_n^*\}$ of G^* are not sufficient for determining $G_{n+1}(\theta)$ since

$$G_{n+1}(\theta) = G^*(\theta - \lambda_{n+1})$$

and λ_{n+1} is unknown. Hence we use the estimate $\lambda_{n+1,n}$ of λ_{n+1} to give

$$G_{n+1,n}(\theta) = \hat{G}_n^*(\theta - \lambda_{n+1,n}) . \tag{3.57}$$

These estimators must now be shown to converge weakly, with probability one, to $G_{n+1}(\theta)$. We have

$$G_{n+1,n}(\theta) - G_{n+1}(\theta) = \hat{G}_n^*(\theta - \lambda_{n+1,n}) - G^*(\theta - \lambda_{n+1})$$

$$= \{\hat{G}_n^*(\theta - \lambda_{n+1,n}) - G^*(\theta - \lambda_{n+1,n})\}$$

$$+ \{G^*(\theta - \lambda_{n+1,n}) - G^*(\theta - \lambda_{n+1})\} \quad . \tag{3.58}$$

Provided G^* is continuous, then (3.54) is satisfied using Theorem 3.11 for the first expression of (3.58) and since $\lambda_{n+1,n} - \lambda_{n+1} \xrightarrow{a.e.} 0$.

As an example, consider the two action problem with $A = \{a_0, a_1\}$ for which (3.6) holds. Defining

$$\Delta_{G_{n+1}}(x) = \int_{\Omega} [L(a,\theta) - L(a_0,\theta)] f(x|\theta) dG_{n+1}(\theta)$$

and

$$\Delta_n(x) = \int_{\Omega} [L(a_1,\theta) - L(a_0,\theta)] f(x|\theta) dG_{n+1,n}(\theta)$$

with $G_{n+1,n}(\theta)$ defined by (3.57), we get that

$$t_n(x) = \begin{cases} 1 & , \text{ if } \Delta_n(x) \geqslant 0 \\ 0 & , \text{ if } \Delta_n(x) < 0 \end{cases}$$

is asymptotically optimal provided that $[L(a,\theta)-L(a_0,\theta)]f(x|\theta)$ is continuous and bounded in θ, by the Helly-Bray theorem.

Unfortunately the method of proof for Theorem 3.10 is non-constructive, and the general problem of determining the sequence $\{\hat{G}_n^*\}$ is not an easy one. Deely and Kruse (1968)

have given a method for finding such a sequence when is a class of distributions defined on a compact subset of the real line. They construct discrete distributions with weights determined by the solution of a linear programming problem. Rutherford and Krutchkoff (1967) give estimates based on moment estimators of a prior assumed to be a member of the Pearson family of distributions.

Extensive work has been made by Maritz (1967, 1968) on smooth empirical Bayes estimators. This involves the approximation of G by a step function $A_k(\theta)$, with k steps of height 1/k at points $u_1, \ldots, u_k$, i.e.,

$$dA_k(\theta) = (u_{j+1} - u_j)/(k-1).du, \; u_j < \theta < u_{j+1}$$

for $j = 1, \ldots, k-1$. Provided the sequence $A_k(\theta)$ is constructed so that $A_k(\theta) \to G(\theta)$ (weakly), then the smooth empirical Bayes approach requires only that estimates are found for the points $u_1, \ldots, u_k$ based on the observations $Y_1, \ldots, Y_n$. While this method offers good numerical results for selected examples, it is analytically intractable.

IV RATES OF CONVERGENCE

(1) Generalised Rates

As for the standard empirical Bayes problem, the investigation of rates of convergence to optimality is an important problem. Consider again the case where the optimal Bayes procedure is given in terms of $f_{G_{n+1}}(x)$ and its

derivative. This often permits rate results to be established on the basis of rates of convergence of estimates $\hat{f}_n^{(1)}(x), \hat{f}_n^{(1)}(x)$ to $\hat{f}_{G_{n+1}}(x)$ and $f_{G_{n+1}}^{(1)}(x)$. In particular consider the kernel function estimators described in section III(3) of this chapter. We are interested in the errors $E|f_{n,n}(y) - f_{G^*}(y)|^{\delta}$, for $\delta > 0$, where $f_{n,n}(y)$ is defined by (3.36).

Lemma 3.11

Assume that the conditions of Theorems 3.5 and 3.6 hold. Suppose that $f_{n,n}(y)$ is defined by (3.36) with $K_1(u)$ satisfying (2.26) and that $f_{n,n}^{(1)}(y)$ is defined by (3.39) with $K_2(u)$ satisfying (2.28). Then, choosing $a_n = O\left(n^{-\frac{1}{2r+1}}\right)$ for some $r \geqslant 3$ we have

$$E|f_{n,n}(y) - f_n(y)|^{\delta} = O\left(n^{-\frac{(2r-3)}{2(2r+1)}\delta}\right), \tag{3.60}$$

$$E|f_{n,n}^{(1)}(y) - f_n^{(1)}(y)| = O\left(n^{-\frac{(2r-5)}{2(2r+1)}\delta}\right), \tag{3.61}$$

where $f_n(y)$ and $f_n^{(1)}(y)$ are given by (3.34) and (3.35) respectively.

Proof

From (3.38)

$$E|f_{n,n}(y) - f_n(y)|^{\delta} \leqslant (2M_1)^{\delta} \sup_{1\leqslant j\leqslant n} \left\{E\left|\frac{Y_{j,n}-Y_j}{a_n^2}\right|^{\delta}\right\}$$

$$\leqslant (2M_1)^{\delta} a_n^{-2\delta} \left(\frac{2\sigma^2}{n} B\right)^{\delta/2}$$

where $B = \sup_i\{1 + (z_i - \bar{z})^2 \Big/ \frac{1}{n}\sum_{i=1}^{n}(z_i-\bar{z})^2\} < \infty$ by assumption

on the $\{z_i\}$. Therefore (3.60) is satisfied. The proof of (3.61) is similar and is therefore omitted.

For examples of asymptotically optimal procedures with a rate of convergence, consider again the one-parameter exponential family (2.20).

(a) Hypothesis Testing

For the two-action problem (3.41) with piecewise linear loss it has been established that the Bayes procedure for this modified problem is $t_{G_{n+1}}(x) = t_{G^*}(y)$ defined by (3.43) with $\alpha_{G_{n+1}}(x) = \alpha(y) = (v^*(y) - \theta_0)f_{G^*}(y) - f^{(1)}_{G^*}(y)$. Theorem 3.8 states that an asymptotically optimal empirical Bayes procedure is

$$t_n(y) = \begin{cases} 1, & \text{if } \alpha_n(y) \leqslant 0 \\ \\ 0, & \text{elsewhere} \end{cases}$$

with $\alpha_n(y)$ given by (3.44). The following theorem gives a rate of convergence for this procedure.

Theorem 3.12

Let θ be a location parameter of the density $f(x|\theta)$ in the family (2.20). Assume $h^{(r)}(x)$ exists for some $r \geqslant 3$. Further assume that for this modified empirical Bayes problem

$$\lambda_{n+1} = \int_{\Omega} \theta dG_{n+1}(\theta) < \infty \quad \text{for all } n. \tag{3.62}$$

Define $\{t_n\}$ by (3.36), (3.39), (3.44) and (3.45). If, for

some δ, $0 < \delta < 2$, and for some $\varepsilon > 0$

$$\int |\alpha(y)|^{1-\delta}(1+\{v^*(y)\}^{\delta})(1 + \{f_{\varepsilon}^{*}(y)\}^{\delta/2})dy < \infty \qquad (3.63)$$

$$\int |\alpha(y)^{1-\delta}(1+\{v^*(y)\}^{\delta})(1+\{q_{\varepsilon}^{(r)}(y)\}^{\delta})dy < \infty \qquad (3.64)$$

where $f_{\varepsilon}^{*}(y) = \sup_{0\leqslant t\leqslant\varepsilon} |f_{G^*}(y+t)|$ and $q_{\varepsilon}^{(r)}(y) = \sup_{0\leqslant t\leqslant\varepsilon} |f_{G^*}^{(r)}(y+t)|$, then with

$$a_n = 0\left(n - \frac{1}{2r+1}\right) \text{ and}$$

$$\int_0^{u_1} u^j K_1(u)du = 0 \quad \int_0^{u_2} u^{j+1} K_2(u)du = 0 \qquad (3.65)$$

for $j = 1, \ldots, r-1$, we obtain $T = \{t_n\}$ is asymptotically optimal in the sense of (3.5) and in fact

$$R_n(T,G_{n+1}) - R(G_{n+1}) = O\left(n^{-\frac{(2r-5)}{2(2r+1)}\delta}\right) . \qquad (3.66)$$

Proof

From Lemma 2.3 we obtain

$$R_n(T,G_{n+1}) - R(G_{n+1}) \leqslant \int_{\mathcal{X}} |\alpha_{G_{n+1}}(x)|^{1-\delta} E|\hat{\alpha}_n(x) - \alpha_{G_{n+1}}(x)|^{\delta}dx$$

$$= \int |\alpha(y)|^{1-\delta}E|\alpha_n(y)-\alpha(y)|^{\delta}dy . \qquad (3.67)$$

Using the C_{δ}-inequality gives

$$E|\alpha_n(y) - \alpha(y)|^{\delta}$$

$$= E\left|(v^*(y)-\theta_o)f_{n,n}(y) - f^{(1)}_{n,n}(y) - (v^*(y) - \theta_o)f_{G^*}(y) + f^{(1)}_{G^*}(y)\right|^{\delta}$$

$$\leq C_\delta E\left|(v^*(y) - \theta_o)(f_{n,n}(y) - f_{G^*}(y))\right|^{\delta}$$

$$+ C_\delta E\left|f^{(1)}_{n,n}(y) - f^{(1)}_{G^*}(y)\right|^{\delta} \quad .$$

Therefore (3.67) gives

$$R_n(T,G_{n+1}) - R(G_{n+1}) \leq A_n + B_n + C_n, \text{ where}$$

$$A_n = C_\delta^2 \int |\alpha(y)|^{1-\delta}\{v^*(y)\}^{\delta}E|f_{n,n}(y)-f_{G^*}(y)|^{\delta}\,dy,$$

$$B_n = C_\delta^2\theta_0^{\delta} \int |\alpha(y)|^{1-\delta}E|f_{n,n}(y) - f_{G^*}(y)|^{\delta}dy \quad ,$$

$$C_n = C_\delta \int |\alpha(y)|^{1-\delta}E|f^{(1)}_{n,n}(y) - f^{(1)}_{G^*}(y)|^{\delta}dy \quad . \qquad (3.68)$$

Johns and Van Ryzin (1972) have shown that with $f_n(y)$ and $f^{(1)}_n(y)$ given by (3.34) and (3.35) respectively

$$E|f_n(y) - f_{G^*}(y)|^{\delta} \leq o\left(n^{-\frac{(r-1)\delta}{2r+1}}\right)[\{f^*_\varepsilon(y)\}^{\delta/2} + \{q^{(r)}_\varepsilon(y)\}^{\delta}]$$

$$E|f^{(1)}_n(y) - f^{(1)}_{G^*}(y)|^{\delta} \leq o\left(n^{-\frac{(r-1)\delta}{2r+1}}\right)[\{f^*_\varepsilon(y)\}^{\delta/2} + \{q^{(r)}_\varepsilon(y)\}^{\delta}]$$

Since

$$E|f_{n,n}(y) - f_{G^*}(y)|^{\delta} \leq C_\delta E|f_{n,n}(y) - f_n(y)|^{\delta}$$

$$+ C_\delta E|f_n(y) - f_{G^*}(y)|^{\delta}$$

we have from Lemma 3.12 that

$$E|f_{n,n}(y) - f_{G^*}(y)|^{\delta} \leq O\left(n^{-\frac{(2r-3)\delta}{2(2r+1)}}\right)[1 + \{f_{\varepsilon}^{*}(y)\}^{\delta/2} + \{q_{\varepsilon}^{(r)}(y)\}^{\delta}], \tag{3.70}$$

since this is the dominating asymptotic rate. Similarly

$$E|f_{n,n}^{(1)}(y) - f_{G^*}(y)|^{\delta} \leq O\left(n^{-\frac{(2r-5)\delta}{2(2r+1)}}\right)[1+\{f_{\varepsilon}^{*}(y)\}^{\delta/2} + \{q_{\varepsilon}^{(r)}(y)\}^{\delta}] . \tag{3.71}$$

Combining (3.70) and (3.63), (3.64) gives

$$A_n,\ B_n = O\left(n^{-\frac{2r-3}{2(2r+1)}\delta}\right) .$$

Also (3.71) with (3.63) and (3.64) yields

$$C_n = O\left(n^{-\frac{(2r-5)}{2(2r+1)}\delta}\right)$$

and since this term dominates asymptotically we have the result.

(b) Estimation

For the exponential family, the Bayes procedure at stage n+1 for the squared error loss estimation problem is $t_{G_{n+1}}(x) = t_{G^*}(y)$ given by (3.46) and (3.47). Consider the empirical Bayes estimator $\hat{t}_n(x) = t_n(y)$, given by (3.48). As before, we must now allow for the case $f_{n,n}(y) = 0$. Let $\eta > 0$ and define

$$\bar{f}_{n,n}(y) = \max\{f_{n,n}(y), \eta\} .$$

This gives the estimator $\bar{t}_n(y)$ $(= \bar{\hat{t}}_n(x))$ where

$$\bar{t}_n(y) = v^*(y) - \frac{f^{(1)}_{n,n}(y)}{\bar{f}_{n,n}(y)} \quad . \tag{3.72}$$

Using Lemma 1.3 the excess risk for the procedure $T = \{t_n\}$ satisfies

$$\begin{aligned} R_n(T,G_{n+1}) - R(G_{n+1}) &= \int_{\mathcal{X}} E|\bar{\hat{t}}_n(x) - t_{G_{n+1}}(x)|^2 f_{G_{n+1}}(x)dx \\ &= \int_{\mathcal{Y}} E|\bar{t}_n(y) - t_{G^*}(y)|^2 f_{G^*}(y)dy \ , \end{aligned} \tag{3.73}$$

(where $x \in \mathcal{X} \equiv x - \lambda_{n+1} \in \mathcal{Y}$). We have the following:

Theorem 3.13

Let θ be a location parameter of $f(x|\theta)$ in the family (2.20). Assume $h^{(r)}$ exists for some $r \geqslant 3$. Let $\bar{t}_n(y)$ be defined by (3.36), (3.39), (3.48) and (3.72). If for some $\varepsilon > 0$

$$\int \left[1 + \left(\frac{f^{(1)}_{G^*}(y)}{f_{G^*}(y)}\right)^2\right]\left[1 + f^*_\varepsilon(y) + (q^{(r)}_\varepsilon(y))^2\right] f_{G^*}(y)dy < \infty, \tag{3.74}$$

where $f^*_\varepsilon(y)$ and $q^{(r)}_\varepsilon(y)$ are given in Theorem 3.14, and if there exists a $\delta > 0$ such that

$$\int_{\{y: f_{G^*}(y) < \eta\}} \{f^{(1)}_{G^*}(y)/f_{G^*}(y)\}^2 f_{G^*}(y)dy \leqslant C_1 \eta^\delta \tag{3.75}$$

for some $C_1 > 0$, then with $a_n = O\left(n^{-\frac{1}{2r+1}}\right)$ and $\eta = n^{-\gamma}$, where $\gamma = (2r-5)/(2r+1)(2+\delta)$ and $K_i(u)$ satisfying (3.65) for $i = 1,2$, the sequence $T = \{\bar{t}_n\}$ is asymptotically optimal in

the sense of (3.5) with

$$R_n(T,G_{n+1}) - R(G_{n+1}) = O(n^{-\gamma\delta}) \quad . \tag{3.76}$$

Proof

First note that with $v^*(y)$ known,

$$E|\bar{t}_n(y) - t_{G^*}(y)|^2 = E\left|\frac{f^{(1)}_{n,n}(y)}{f_{n,n}(y)} - \frac{f^{(1)}_{G^*}(y)}{f_{G^*}(y)}\right|^2$$

$$\leq \eta^{-2}E\left|f^{(1)}_{n,n}(y) - \bar{f}_{n,n}(y)\cdot\frac{f^{(1)}_{G^*}(y)}{f_{G^*}(y)}\right|^2$$

$$= \eta^{-2}\{2E|f^{(1)}_{n,n}(y) - f^{(1)}_{G^*}(y)|^2$$

$$+ 2\left|\frac{f^{(1)}_{G^*}(y)}{f_{G^*}(y)}\right|^2 E|\bar{f}_{n,n}(y) - f_{G^*}(y)|^2\} \quad .$$

Further, since

$$E|\bar{f}_{n,n}(y) - f_{G^*}(y)|^2$$

$$\leq \eta^{-2}\{y:f_{G^*}(y) < \eta\} + E|f_{n,n}(y) - f_{G^*}(y)|^2$$

we have, from (3.73),

$$R_n(T,G_{n+1}) - R(G_{n+1}) \leq A_n + B_n + C_n$$

where

$$A_n = 4\eta^{-2}\int E|f^{(1)}_{n,n}(y) - f^{(1)}_{G^*}(y)|^2 f_{G^*}(y)dy,$$

$$B_n = 8\eta^{-2} \int E|f_{n,n}(y) - f_{G^*}(y)|^2 \cdot \left|\frac{f_{G^*}^{(1)}(y)}{f_{G^*}(y)}\right|^2 f_{G^*}(y)dy$$

$$C_n = 8\eta^{-2} \int_{\{y: f_{G^*}(y) < \eta\}} \left|\frac{f_{G^*}^{(1)}(y)}{f_{G^*}(y)}\right|^2 f_{G^*}(y)dy \quad .$$

Thus, using the results of Theorem 3.12, namely (3.70), (3.71), together with (3.74) and (3.75) we get

$$A_n = O\left(n^{\frac{(2r-5)\delta}{2(2r+1)}} \cdot \eta^{-2}\right)$$

$$B_n = O\left(n^{-\frac{(2r-3)\delta}{2(2r+1)}} \eta^{-2}\right)$$

$$C_n = O(n^{-\gamma\delta}).$$

Thus we get,

$$R_n(T, G_{n+1}) - R(G_{n+1}) = O\left(n^{-\frac{\delta}{2+\delta}\frac{(2r-5)}{2r+1}}\right),$$

since this is the dominating asymptotic order.

The proofs of the above two theorems rely on the rate results produced by Johns and Van Ryzin (1972) and Lin (1975). It would be possible to take the theorems and apply them to particular densities on the exponential family in order to reduce the unintuitive conditions (3.63), (3.64) and (3.74), (3.75) to moment restrictions on the prior distribution $G_{n+1}(\theta) = G^*(\theta - \lambda_{n+1})$. However as explained in Chapter 2, even these conditions do not permit easy evaluation of the bounds on the excess risk. The rate results above, while allowing for a theoretical rate of convergence arbitrarily

close to $O(n^{-1})$ do not give good small sample properties.

Extension of the Model

It has been assumed throughout this chapter that the variance of θ_i remains constant. This was done to ensure that the unconditional variance of X_i remains fixed to satisfy an assumption of the usual linear regression model. In fact we may assume that this variance is not constant provided it is a known value. This does not affect the estimator for λ_i, but it does affect the expression for the expected squared error. In fact if

$$\mathrm{Var}(X_i) = \sigma_i^2$$

then with estimators (3.17), (3.18), (3.19) for the prior mean λ_i based on $X_1, \ldots, X_k$ we have similar to Lemma 3.3

$$E|\lambda_{i,k} - \lambda_i|^2 = E[\bar{X} - \bar{\lambda} + (z_i-\bar{z})(a_k-a)]^2$$

$$\leqslant 2E|\bar{X}-\bar{\lambda}|^2 + (z_i - \bar{z})^2 E(a_k-a)^2$$

$$= 2\left\{\frac{1}{k^2}\sum_{i=1}^{k}\sigma_i^2 + (z_i - \bar{z})^2 \frac{\sum_{i=1}^{k}\sigma_i^2(z_i-\bar{z})^2}{\sum_{i=1}^{k}(z_i-\bar{z})^2}\right\} \qquad (3.77)$$

If it is assumed that σ_i^2 is bounded above by σ^2 say then (3.20) will be satisfied and the conditions governing convergence will also apply here.

The general methods developed in this chapter depend upon the kernel type estimators of a probability density function.

Other methods are available (see Wegman (1972)), but as with the kernel functions, the accuracy of such estimators for small samples is often difficult to determine. Consequently the assumption of a parametric form for the prior distributions takes on some practical significance.

(2) Parametric Examples

Consider again the case where it is assumed that $G_n(\theta)$ is a member of a given parametric family.

(a) Normal-Normal

Suppose $f(x|\theta)$ is $N(\theta,\beta^2)$ and $G_{n+1}(\theta)$ is $N(\lambda_{n+1}, \sigma^2)$. We are interested in the squared error loss estimation of $\theta = \theta_{n+1}$. The unconditional distribution of $X = X_{n+1}$ is $f_{G_{n+1}}(x)$ which has a $N(\lambda_{n+1}, \sigma^2 + \beta^2)$ density. The Bayes procedure is

$$t_{G_{n+1}}(x) = \frac{\beta^2\lambda_{n+1} + \sigma^2 x}{\sigma^2 + \beta^2} \quad . \tag{3.78}$$

Consequently using the estimate $\lambda_{n+1,n+1}$ for λ_{n+1} based on $X_1, \ldots, X_n$ and $X_{n+1} = X$, we define

$$t_n(x) = \frac{\beta^2\lambda_{n+1,n+1} + \sigma^2 x}{\sigma^2 + \beta^2} \quad . \tag{3.79}$$

and by Lemma 1.3, the excess risk at stage n+1 for the procedure $T = \{t_n\}$ is

$$R(T,G_{n+1}) - R(G_{n+1}) = \int_{\mathcal{X}} E|t_n(x) - t_{G_{n+1}}(x)|^2 f_{G_{n+1}}(x)dx$$

$$= \left\{\frac{\beta^2}{\sigma^2+\beta^2}\right\}^2 E|\lambda_{n+1,n+1} - \lambda_{n+1}|^2$$

$$\leqslant \left\{\frac{\beta^2}{\sigma^2\ \beta^2}\right\} \frac{2(\sigma^2+\beta^2)}{n+1} \left\{1 + \frac{(z_{n+1} - \bar{z})^2}{\frac{1}{n+1} \sum_{i=1}^{n+1} (z_i-\bar{z})^2}\right\},$$

by (3.20),

$$\leqslant \frac{2\beta^4}{\sigma^2+\beta^2} \cdot B \cdot \frac{1}{n+1} \tag{3.80}$$

where $B = \sup_i \{1 + (z_i-\bar{z})^2 / \frac{1}{n+1} \sum_{i=1}^{n+1} (z_i - \bar{z})^2\} < \infty$.
That is we have a convergence rate of $O(\frac{1}{n+1})$ and exact error bounds are easy to determine.

We now note that if we assume a parametric form for $G_{n+1}(\theta)$ then since the Bayes procedure is often expressed in terms of the a priori parameters it may be possible to remove the condition that $f(x|\theta)$ and $g_{n+1}(\theta) = G'_{n+1}(\theta)$ are continuous location parameter densities. The direct estimation of the a priori distribution parameters may be sufficient to estimate the Bayes procedure. To illustrate this, consider the following.

(b) Poisson-Gamma

Suppose the observations are conditionally drawn from the density

$$f(x|\theta) = \frac{\theta^x e^{-\theta}}{x!} \qquad x = 0,1,2, \ldots ,$$

and that the a priori distributions $G_n(\theta)$ have the density

$$g_n(\theta) = \frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n)} \theta^{\alpha_n-1} e^{-\beta_n\theta} \quad , \quad \theta > 0,$$

where $\{\beta_n\}$ is known. It is further assumed that

$$\alpha_n = az_n + b$$

where $\{z_n\}$ is known, a,b unknown. Similar to Example 1 in Chapter II, part II we have for each i

$$E[X_i] = E_{G_i}[\theta] = \alpha_i/\beta_i, \text{ for all } i,$$

and for the problem of squared error loss estimation of $\theta_{n+1} = \theta$, the Bayes procedure is the posterior mean

$$t_{G_{n+1}}(x) = \frac{\alpha_{n+1} + x}{\beta_{n+1} + 1} . \tag{3.81}$$

Consequently we may provide an empirical Bayes estimation procedure upon estimating α_{n+1}. Consider the estimator based on $X_1, \ldots, X_n, X_{n+1} = X$

$$\alpha_{n+1,n+1} = a_{n+1} z_{n+1} + b_{n+1}$$

where a_{n+1}, b_{n+1} are given by (3.17) and (3.18) respectively. This gives

$$t_n(x) = \frac{\alpha_{n+1,n+1} + x}{\beta_{n+1} + 1} .$$

Furthermore, similar to (2.40) we have

$$0 \leqslant R_n(T,G_{n+1}) - R(G_{n+1}) = E|t_n - t_{G_{n+1}}|^2$$

$$= \frac{1}{(\beta_{n+1} + 1)^2} E|\alpha_{n+1,n+1} - \alpha_{n+1}|^2 \quad . \tag{3.82}$$

Also, since $\mathrm{Var}(X_i) = \alpha_i(\beta_i+1)/\beta_i^2$, we use (3.77) for $k = n+1$ and the r.h.s. of (3.82) is bounded by

$$\frac{2}{(\beta_{n+1}+1)^2}\left|\frac{1}{(n+1)^2}\sum_{i=1}^{n+1}\frac{\alpha_i(\beta_i+1)}{\beta_i^2} + \frac{(z_i-\bar{z})^2\sum_{i=1}^{n+1}\frac{\alpha_i(\beta_i+1)(z_i-\bar{z})^2}{\beta_i^2}}{\sum_{i=1}^{n+1}(z_i-\bar{z})^2}\right| . \tag{3.83}$$

Provided $\dfrac{\alpha_i(\beta_i+1)}{\beta_i^2} \leqslant A < \infty$ (i.e. $\mathrm{Var}(X_i)$ is bounded), (3.83) gives a convergence rate of $O((n+1)^{-1})$.

CHAPTER IV

A SEQUENTIAL SELECTION PROBLEM

I. INTRODUCTION

Suppose we have k populations $\pi_1, \ldots, \pi_k$. From each π_j we observe a random variable $X^{(j)}$ whose conditional density is $f(x|\theta^{(j)})$, where $\theta^{(j)} \in \Omega$, $j = 1, \ldots, k$. On the basis of the observation

$$\underline{X} = (X^{(1)}, \ldots, X^{(k)})$$

it is desired to select the population associated with the largest parameter value $\theta^{[k]}$. It is assumed that, conditional upon $\underline{\theta} = (\theta^{(1)}, \ldots, \theta^{(k)})$, the $X^{(j)}$ are independent. The action space in this case is $A = \{a_1, \ldots, a_k\}$, where a_i means 'say $\theta^{(i)}$ is largest'.

This problem of selecting the 'best' of several population categories is important to many practical applications, some of which are discussed in Dunnett (1960). In the Bayesian approach to this problem it is assumed that each of the parameters $\theta^{(i)}$ has the a priori distribution $G^{(i)}(\theta^{(i)})$. We therefore define the a priori distribution $G(\underline{\theta})$ over Ω^k as

$$G(\underline{\theta}) = \prod_{i=1}^{k} G^{(i)}(\theta^{(i)}) \quad . \tag{4.1}$$

(1) The Bayes Procedure Under Linear Loss

We shall consider the linear loss function

$$L(a_i, \underline{\theta}) = \theta^{[k]} - \theta^{(i)}, \text{ for } i = 1, \ldots, k \quad . \tag{4.2}$$

Recalling the notation of Chapter I, we have by (1.4)

$$\phi_G(a_i, \underline{x})$$

$$= \int_{\Omega^k} \theta^{[k]} f(\underline{x}|\underline{\theta}) dG(\underline{\theta}) - \int_{\Omega^k} \theta^{(i)} f(\underline{x}|\underline{\theta}) dG(\underline{\theta}) \quad . \tag{4.3}$$

The Bayes procedure is, from (1.7)

$$t_G(\underline{x}) = a_j \text{ where } j \text{ is any integer } 1, \ldots, k$$

such that

$$\phi_G(a_j, \underline{x}) = \min_{1 \leqslant i \leqslant k} \phi_G(a_i, \underline{x}) \quad . \tag{4.4}$$

Since the first integral of (4.3) is independent of 'i', the Bayes procedure is now $t_G(\underline{x}) = a_j$ where j is any integer 1, ... , k such that

$$\int_{\Omega^k} \theta^{(j)} f(\underline{x}|\underline{\theta}) dG(\underline{\theta}) = \max_{1 \leqslant i \leqslant k} \int_{\Omega^k} \theta^{(i)} f(\underline{x}|\underline{\theta}) dG(\underline{\theta}) \quad . \tag{4.5}$$

To examine the quantities $\int_{\Omega^k} \theta^{(i)} f(\underline{x}|\underline{\theta}) dG(\underline{\theta})$ more closely we assume that each $G^{(i)}(\theta^{(i)})$ has a density function $g^{(i)}(\theta^{(i)})$. We have

$$\int_{\Omega^k} \theta^{(i)} f(\underline{x}|\underline{\theta}) dG(\underline{\theta})$$

$$= \left\{\int_{\Omega} \theta^{(i)} f(x^{(i)}|\theta^{(i)}) dG^{(i)}(\theta^{(i)})\right\}\left\{\prod_{\substack{j=1 \\ j\neq i}}^{k} \int_{\Omega} f(x^{(j)}|\theta^{(j)}) dG^{(j)}(\theta^{(j)})\right\} . \tag{4.6}$$

The first integral on the r.h.s. of (4.6) is

$$\int_{\Omega} \theta^{(i)} f(x^{(i)}|\theta^{(i)}) g^{(i)}(\theta^{(i)}) d\theta^{(i)}$$

$$= \int_{\Omega} \theta^{(i)} g^{(i)}(\theta^{(i)}|x^{(i)}) f_G^{(i)}(x^{(i)}) d\theta^{(i)} \tag{4.7}$$

where

$$f_G^{(i)}(x^{(i)}) = \int_{\Omega} f(\theta^{(i)}|x^{(i)}) g^{(i)}(\theta^{(i)}) d\theta^{(i)} \tag{4.8}$$

is the unconditional density of $X^{(i)}$, and

$$g^{(i)}(\theta^{(i)}|x^{(i)}) = \frac{f(x^{(i)}|\theta^{(i)}) g^{(i)}(\theta^{(i)})}{f_G^{(i)}(x^{(i)})} \tag{4.9}$$

is the posterior density of $\theta^{(i)}$ given the observation $X^{(i)} = x^{(i)}$. Therefore upon setting

$$f_G(\underline{x}) = \prod_{i=1}^{k} f_G^{(i)}(x^{(i)}) \tag{4.10}$$

we obtain from (4.6) and (4.7)

$$\int_{\Omega^k} \theta^{(i)} f(\underline{x}|\underline{\theta}) dG(\underline{\theta}) = \left\{\int_{\Omega} \theta^{(i)} g^{(i)}(\theta^{(i)}|x^{(i)}) d\theta^{(i)}\right\} \cdot f_G(\underline{x})$$

$$= E[\theta^{(i)}|x^{(i)}] \cdot f_G(\underline{x}) . \tag{4.11}$$

This proves the following:

Theorem 4.1

From each of k populations, observe a random variable which has a density function $f(x|\theta^{(i)})$ for i = 1, ... , k. If $\theta^{(i)}$ is distributed according to the density $g^{(i)}(\theta^{(i)})$, then the Bayes procedure for selecting the largest population parameter under the linear loss function (4.2) is given by

$$t_G(\underline{x}) = a_j \text{ where } j \text{ is any integer } 1, \ldots, k$$

such that

$$E[\theta^{(j)}|x^{(j)}] = \max_{1 \leq i \leq k} E[\theta^{(i)}|x^{(i)}] \qquad (4.12)$$

where $E[\theta^{(i)}|x^{(i)}]$ is the a posteriori mean of the ith population given the observation $X^{(i)} = x^{(i)}$.

To illustrate the above result consider the Normal-Normal example. That is, $f(x^{(i)}|\theta^{(i)})$ has the density of a $N(\theta^{(i)}, \sigma_i^2)$ random variable and the a priori density of $\theta^{(i)}$ is that of a $N(\lambda^{(i)}, \beta_i^2)$ random variable. The posterior distribution thus has the density (see Deely (1965))

$$g^{(i)}(\theta^{(i)}|x^{(i)}) \sim N\left(\frac{\beta_i^2 x^{(i)} + \sigma_i^2 \lambda^{(i)}}{\beta_i^2 + \sigma_i^2}, \frac{\beta_i^2 \sigma_i^2}{\beta_i^2 + \sigma_i^2}\right) .$$

Therefore, the Bayes procedure for our selection problem is

$$t_G(\underline{x}) = a_j \text{ where } j \text{ is any integer } 1, \ldots, k$$

such that

$$\frac{\beta_j^2 x^{(j)} + \sigma_j^2 \lambda^{(j)}}{\beta_j^2 + \sigma_j^2} = \max_{1 \leqslant i \leqslant k} \frac{\beta_i^2 x^{(i)} + \sigma_i^2 \lambda^{(i)}}{\beta_i^2 + \sigma_i^2} \tag{4.13}$$

For simplicity of notation we have assumed that the observations $X^{(i)}$ are one dimensional. In fact all of the above results generalise simply to the case where each $X^{(i)}$ is a vector of $m(\geqslant 1)$ observations of $f(x|\theta^{(i)})$.

(2) The Bayes Procedure Under 0-1 Loss

Consider now the selection problem as above under the loss structure

$$L(a_i, \underline{\theta}) = \begin{cases} 0 & , \text{ if } \theta^{(i)} = \theta^{[k]} \\ 1 & , \text{ otherwise} \end{cases} . \tag{4.14}$$

From (1.4)

$$\phi_G(a_i, \underline{x}) = \int_{\Omega^k} L(a_i, \underline{\theta}) f(\underline{x}|\underline{\theta}) g(\underline{\theta}) d\underline{\theta}$$

$$= \int_{\Omega^k} L(a_i, \underline{\theta}) \prod_{i=1}^{k} f(x^{(i)}|\theta^{(i)}) g^{(i)}(\theta^{(i)}) d\underline{\theta}$$

$$= \int_{\Omega^k} L(a_i, \underline{\theta}) \prod_{i=1}^{k} g^{(i)}(\theta^{(i)}|x^{(i)}) f_G^{(i)}(x^{(i)}) d\underline{\theta}$$

from (4.9),

$$= f_G(\underline{x}) \cdot \int_{\Omega^k} L(a_i, \underline{\theta}) g(\underline{\theta}|\underline{x}) d\theta \tag{4.15}$$

from (4.10), with

$g(\underline{\theta}|\underline{x}) = \prod_{i=1}^{k} g^{(i)}(\theta^{(i)}|x^{(i)})$. Using the 0-1 loss, (4.15) becomes

$$f_G(\underline{x}) \cdot \int_{\{\underline{\theta}:\theta^{(i)} \neq \theta^{[k]}\}} g(\underline{\theta}|\underline{x})\,d\underline{\theta}$$

$$= f_G(\underline{x}) \cdot P(\theta^{(i)} \neq \theta^{[k]}|\underline{x})$$

$$= f_G(\underline{x}) \cdot \{1-P(\theta^{(i)} = \theta^{[k]}|\underline{x})\} \quad . \tag{4.16}$$

This gives the following result.

Theorem 4.2

Assume the conditions of Theorem 4.1 hold. The Bayes procedure for selecting the best population parameter under the loss structure (4.14) is to select the population with the largest a posteriori probability of being best; i.e.

$t_G(\underline{x}) = a_j$ where j is any integer 1, ... , k

such that

$$P(\theta^{(j)} = \theta^{[k]}|\underline{x}) = \max_{1 \leq i \leq k} P(\theta^{(i)} = \theta^{[k]}|\underline{x}) \quad . \tag{4.17}$$

Consider again the Normal-Normal case.

$$f(x^{(i)}|\theta^{(i)}) \sim N(\theta^{(i)},\sigma^2),\ g^{(i)}(\theta^{(i)}) \sim N(\lambda^{(1)},\sigma_\theta^2)$$

The posterior distribution of $\theta^{(i)}$ given $x^{(i)}$ is therefore

$$N\left(\frac{\sigma_0^2 x^{(i)} + \sigma^2 \lambda^{(i)}}{\sigma^2 + \sigma_0^2}, \frac{\sigma^2\sigma_0^2}{\sigma^2+\sigma_0^2}\right)$$

$$= N(\eta^{(i)}, \delta^2) \text{ , say } .$$

We require the posterior probability

$$P(\theta^{(j)} = \theta^{[k]} | \underline{x}) = P(\theta^{(i)} \leqslant \theta^{(j)} \text{ for all } i \neq j | \underline{x})$$

$$= P\left(\bigcap_{\substack{i=1 \\ i\neq j}}^{k} \frac{\theta^{(i)} - \nu^{(i)}}{\delta} \leqslant \frac{\theta^{(j)} - \eta^{(j)}}{\delta} + \frac{\eta^{(j)} - \eta^{(i)}}{\delta} \,\middle|\, \underline{x}\right)$$

$$= P\left(\bigcap_{\substack{i=1 \\ i\neq j}}^{k} z^{(i)} \leqslant z^{(j)} + \frac{\eta^{(j)} - \eta^{(i)}}{\delta} \,\middle|\, \underline{x}\right)$$

where $z^{(i)}$, $i = 1, \ldots, k$, are i.i.d. $N(0,1)$ random variables, and by Chung (1968)

$$= \int_{-\infty}^{\infty} \prod_{\substack{i=1 \\ i\neq j}}^{k} \Phi\left(z + \frac{\eta^{(j)} - \eta^{(i)}}{\delta}\right) \phi(z)\,dz \qquad (4.18)$$

where ϕ and Φ are respectively the p.d.f and c.d.f. of an $N(0,1)$ random variable. The Bayes procedure of (4.17) requires that we choose the value of j for which (4.18) is largest. Clearly this is done by choice of the population with largest $\eta^{(i)}$. That is the Bayes procedure for the 0-1 loss problem is to choose the population with the largest posterior mean. Note however that this result is valid only for the Normal-Normal case with equal population variances.

II THE EMPIRICAL BAYES APPROACH WITH A CHANGING PRIOR

(1) General Case

Consider the situation in which each of the populations π_i represents a consignment of a machine part. It is desired on the basis of a sample to choose the population with the smallest fraction defective. Due to variations in the manufacturing process, the fraction defective $\theta^{(i)}$ of consignment i (i = 1, ... , k) is distributed according to $G_1^{(i)}(\theta^{(i)})$ at some time t_1 but that at another time t_2, it is distributed according to $G_2^{(i)}(\theta^{(i)})$, perhaps due to deterioration of the manufacturing process (or indeed, improvement in the process). The variation in these distributions over time need not be the same for each manufacturer. The problem thus is a sequential choice for the population with smallest fraction defective.

If the a priori distributions $G_n^{(i)}(\theta^{(i)})$, i=1, ... , k, were known for each n, then the problem is the Bayesian one of the previous section. In particular suppose all a priori distributions have a derivative and that $G_n(\underline{\theta}) = \prod_{i=1}^{k} G_n^{(i)}(\theta^{(i)})$. The Bayes procedure, at time n for selecting the best population (with the largest fraction non-defective) under the linear loss (4.2) based on the observation $\underline{X}_n=(X_n^{(1)},\ldots,X_n^{(k)})$ taken on $f_{G_n}(\underline{x})$, is

$$t_{G_n}(x) = a_j \text{ where } j \text{ is any integer } 1, \ldots, k \text{ such that}$$

$$\int_{\Omega^{(j)}} \theta^{(j)} g_n^{(j)}(\theta^{(j)}|x^{(j)})\, d\theta^{(j)}$$

$$= \max_{1\leqslant i\leqslant k} \int_{\Omega} \theta^{(i)} g_n^{(i)}(\theta^{(i)}|x^{(i)})\, d\theta^{(i)} \quad , \tag{4.19}$$

where $g_n^{(i)}(\theta^{(i)}|x^{(i)})$ is the a posteriori density of $\theta^{(i)}$ given $X_n^{(i)} = x^{(i)}$, as given by (4.9).

Suppose however that only existence of the distributions $\{G_n\}$ is postulated, but that previous history in the form of the independent observations $(\underline{X}_1,\underline{\theta}_1)$, $(\underline{X}_2,\underline{\theta}_2)$, ... , $(\underline{X}_n,\underline{\theta}_n)$, where $\underline{X}_r = (X_r^{(1)}, \ldots, X_r^{(k)})$ is observed on the conditional density

$$f(\underline{x}|\underline{\theta}_r) = \prod_{i=1}^{k} f(x^{(i)}|\theta_r^{(i)})$$

and for each $i = 1, \ldots, k$, $\theta_r^{(i)}$ has the a priori distribution $G_r^{(i)}(\theta^{(i)})$. Thus, if we are at the (n+1)st stage of observation, the prior is G_{n+1} and on the basis of $\underline{X}_1, \ldots, \underline{X}_n$ it is desired to construct an estimator $t_n(\underline{x})$ for the Bayes procedure $t_{G_{n+1}}(\underline{x})$ given by (4.19). Adopting the convention of Chapter III we require that this procedure is asymptotically optimal in the sense of (3.5).

Recalling (4.3) we define

$$\phi_{G_{n+1}}(a_i,\underline{x}) = \int_{\Omega^k} \theta^{[k]} f(\underline{x}|\underline{\theta}) dG_{n+1}(\theta)$$

$$- \int_{\Omega} \theta^{(i)} f(\underline{x}|\underline{\theta}) dG_{n+1}(\underline{\theta})$$

and

$$\Delta_{G_{n+1}}(a_i,\underline{x}) = \phi_{G_{n+1}}(a_i,\underline{x}) - \phi_{G_{n+1}}(a_1,\underline{x})$$

$$= \int_{\Omega^k} \theta^{(1)} f(\underline{x}|\underline{\theta}) dG_{n+1}(\underline{\theta}) - \int_{\Omega^k} \theta^{(i)} f(\underline{x}|\underline{\theta}) dG_{n+1}(\underline{\theta})$$

$$= \left\{ \int_{\Omega} \theta^{(1)} g_{n+1}^{(1)}(\theta^{(1)}|x^{(1)}) d\theta^{(1)} \right.$$

$$\left. - \int_{\Omega} \theta^{(i)} g_{n+1}^{(i)}(\theta^{(i)}|x^{(i)}) d\theta^{(i)} \right\} \cdot f_{G_{n+1}}(\underline{x}) \quad , \qquad (4.20)$$

by (4.11). The Bayes procedure at stage n+1 may now be written

$$t_{G_{n+1}}(\underline{x}) = a_j \quad \text{where } j \text{ is any integer } 1, \ldots, k$$

such that

$$\Delta_{G_{n+1}}(a_j, \underline{x}) = \min_{1 \leqslant i \leqslant k} \Delta_{G_{n+1}}(a_i, \underline{x}) \quad . \qquad (4.21)$$

The following theorem is useful.

Theorem 4.3

Suppose the distributions $G_n(\underline{\theta})$ are such that for all $i=1, \ldots, k$ and for all n

$$\int_{\Omega} |\theta^{(i)}| dG_n^{(i)}(\theta^{(i)}) < \infty \quad . \qquad (4.22)$$

Let $\Delta_n(a_i, \underline{x}) = \Delta_n(a_i; \underline{x}_1, \ldots, \underline{x}_n; \underline{x})$ be such that for each $i = 1, \ldots, k$

$$\Delta_n(a_i, \underline{x}) - \Delta_{G_{n+1}}(a_i, \underline{x}) \xrightarrow{p} 0, \text{ as } n \to \infty \qquad (4.23)$$

then defining the empirical procedure

$$t_n(\underline{x}) = a_j \text{ where } j \text{ is any integer } 1, \ldots, k \text{ such that}$$

$$\Delta_n(a_j,\underline{x}) = \min_{1\leqslant i\leqslant k} \Delta_n(a_i,\underline{x}) \tag{4.24}$$

we have that, for the problem of selecting the population with the largest population parameter under the linear loss (4.2), $T = \{t_n\}$ is asymptotically optimal in the sense of (3.5).

Proof

First note that

$$\int_{\Omega^k} L(a_i,\underline{\theta})\,dG_n(\underline{\theta}) = \int_{\Omega^k} (\theta^{[k]} - \theta^{(i)})\,dG_n(\underline{\theta})$$

$$= \int_{\Omega^k} (\theta^{[k]} - \theta^{(i)}) \prod_{j=1}^{k} g_n^{(j)}(\theta^{(i)})\,d\underline{\theta}$$

$$\leqslant \int_{\Omega^k} \left(\sum_{j=1}^{k} |\theta^{(j)}| - \theta^{(i)} \right) \prod_{j=1}^{k} g_n^{(j)}(\theta^{(j)})\,d\underline{\theta}$$

$$= \sum_{j=1}^{k} \int_{\Omega} |\theta^{(j)}| g_n^{(j)}(\theta^{(j)})\,d\theta^{(j)} - \int_{\Omega} \theta^{(i)} g_n^{(i)}(\theta^{(i)})\,d\theta^{(i)}$$

$< \infty$, by (4.22) .

Therefore, if $L(\theta) = \max_{1\leqslant i\leqslant k} L(a_i,\underline{\theta})$ we have satisfied (3.6). It is now necessary to show only that

$$\Delta_n(t_n(\underline{x}),\underline{x}) - \Delta_{G_{n+1}}(t_{G_{n+1}}(\underline{x}),\underline{x}) \xrightarrow{p} 0 \quad .$$

This is achieved using a proof analagous to that of Theorem 1.2. Thus by Theorem 3.1 we have asymptotic optimality.

From (4.20) we see that in order to construct consistent estimators for $\Delta_{G_{n+1}}(a_i, \underline{x})$ it is necessary to find consistent estimators for

(i) The posterior means

$$E_{G_{n+1}}[\theta^{(i)}|x^{(i)}] = \int_{\Omega} \theta^{(i)} g_{n+1}(\theta^{(i)}|x^{(i)}) d\theta^{(i)}$$

and

(ii) The unconditional densities $f^{(i)}_{G_{n+1}}(x^{(i)})$.

We will assume that

(a) The conditional density function $f(\underline{x}|\underline{\theta})$ is such that the posterior means may be expressed as functions of the density $f^{(i)}_{G_{n+1}}(x^{(i)})$ and its derivatives, then all we require is consistent estimators for these in order to achieve asymptotic optimality. Examples of densities which satisfy this requirement are

(i) $f(x|\theta) = h(x)\theta^x \beta(\theta) \qquad \theta \in \Omega,\ x \in \mathcal{X},$

for which

$$E(\theta|X=x] = \int_{\Omega} \theta^{x+1} h(x)\beta(\theta) dG(\theta)$$

$$= \frac{h(x)}{h(x+1)} \cdot \frac{f_G(x+1)}{f_G(x)} ,$$

and (ii) the continuous one-parameter exponential density

$$f(x|\theta) = e^{-\theta x} h(x)\beta(\theta) , \qquad \theta \in \Omega , x \in \mathcal{X} ,$$

with

$$E[\theta|X=x] = \frac{h'(x)}{h(x)} f_G(x) - f_G'(x) .$$

(b) For each i = 1, ... , k $\theta^{(i)}$ is a location parameter for the density $f(x^{(i)}|\theta^{(i)})$ and further that the a priori distributions $G_n(\theta^{(i)})$ have a location parameter density for which

$$g_n(\theta^{(i)}) = g^*(\theta^{(i)} - \lambda_n^{(i)}) \tag{4.25}$$

where

$$\lambda_n^{(i)} = \int_{\Omega} \theta^{(i)} dG_n^{(i)}(\theta^{(i)})$$

is the a priori mean of parameter i at stage n.

(c) The a priori means change in the following fashion:

$$\lambda_j^{(i)} = a^{(i)} Z_j^{(i)} + b^{(i)} \tag{4.26}$$

where $\{Z_j^{(i)}\}_{j=1}^{\infty}$ is a known sequence for each i = 1, ... , k.

Define the random variables

$$\underline{Y}_j = \underline{X}_j - \underline{\lambda}_j \tag{4.27}$$

where $\underline{\lambda}_j = (\lambda_j^{(1)}, \ldots, \lambda_j^{(k)})$. We have for the jth observation,

$$f_{G_j}(\underline{x}_j) = \prod_{i=1}^{k} f_{G_j}^{(i)}(x_j^{(i)})$$

$$= \prod_{i=1}^{k} f_{G^*}^{(i)}(y_j^{(i)}) , \qquad \text{from (3.14)}$$

$$= f_{G^*}(\underline{y}_j) \quad , \quad \text{for all } j. \tag{4.28}$$

Thus the $\underline{Y}_j$ are independent and identically distributed random variables with the density function $f_{G^*}(\cdot)$.

If we knew the sequence $\{\underline{\lambda}_n\}$ then we could form the i.i.d. random variables $\underline{Y}_1, \ldots, \underline{Y}_n, \underline{Y}_{n+1}$ and use these to construct estimators for the density $f_{G^*}(\underline{y})$ and thus obtain asymptotically optimal rules in the usual empirical Bayes fashion. Note however that for examples in which the Bayes procedure is a function of the a priori mean even this is not necessary, since we then know the exact Bayes procedure.

However if both of the parameters $a^{(i)}$, $b^{(i)}$ are unknown, but we have the previous observations $\underline{X}_1, \ldots, \underline{X}_n$ then we may, for each fixed $i = 1, \ldots, k$, use the sequence of ith components $X_1^{(i)}, \ldots, X_n^{(i)}$ to estimate $a^{(i)}$, $b^{(i)}$ as in Chapter III. In fact we have the estimators

$$a_n^{(i)} = \frac{\frac{1}{n}\sum_{j=1}^{n}(x_j^{(i)} - \bar{x}^{(i)})(z_j^{(i)} - \bar{z}^{(i)})}{\frac{1}{n}\sum_{j=1}^{n}(z_j^{(i)} - \bar{z}^{(i)})^2} \quad , \tag{4.29}$$

$$b_n^{(i)} = \bar{x}^{(i)} - a_n^{(i)}\bar{z}^{(i)} \quad , \tag{4.30}$$

for $i = 1, \ldots, k$. This gives $\underline{\lambda}_{j,n} = (\lambda_{j,n}^{(1)}, \ldots, \lambda_{j,n}^{(k)})$, where

$$\lambda_{j,n}^{(i)} = a_n^{(i)} z_j^{(i)} + b_n^{(i)} \quad . \tag{4.31}$$

For each sequence $\{z_j^{(i)}\}_{j=1}^{\infty}$ we assume that (3.21) is valid. Consider, for each $j = 1, \ldots, n+1$, the k-vector random

variables

$$\underline{Y}_{j,n} = \underline{X}_j - \underline{\lambda}_{j,n} \quad . \tag{4.32}$$

Using these random variables we may now at stage n+1, upon observing $\underline{X}_{n+1} = \underline{x}$, construct consistent estimators $f_{n,n}(\underline{y})$ and $f'_{n,n}(\underline{y})$ for the 'present' unconditional density $f_{G_{n+1}}(\underline{x}) = f_{G^*}(\underline{y})$ and its derivative.

Suppose now that for each i = 1, ... , k, $f(x^{(i)}|\theta^{(i)})$ is a member of the exponential family (3.11). In particular

$$f(x^{(i)}|\theta^{(i)}) = h^{(i)}(x^{(i)})\exp\{-\theta^{(i)}x^{(i)}\}\beta^{(i)}(\theta^{(i)}) \quad . \tag{4.33}$$

Using Theorem 4.1 the Bayes procedure for selecting the best population under the linear loss (4.2) when the a priori distribution is

$$G_{n+1}(\underline{\theta}) = \prod_{i=1}^{k} G_{n+1}^{(i)}(\theta^{(i)}) \quad ,$$

is given by

$$t_{G_{n+1}}(\underline{x}) = a_j \text{ where } j \text{ is any integer } 1, \ldots , k$$

such that

$$\frac{h'^{(j)}(x^{(j)})}{h^{(j)}(x^{(j)})} f_{G_{n+1}}^{(j)}(x^{(j)}) - f_{G_{n+1}}'^{(j)}(x^{(j)})$$
$$= \max_{1\leq i\leq k}\left\{\frac{h'^{(i)}(x^{(i)})}{h^{(i)}(x^{(i)})} f_{G_{n+1}}^{(i)}(x^{(i)}) - f_{G_{n+1}}'^{(i)}(x^{(i)})\right\} \tag{4.34}$$

[Here f' denotes the first derivative].

Thus, with $h^{(i)}(x^{(i)})$ known, we require estimators for the unconditional density $f^{(i)}_{G_{n+1}}(x^{(i)})$ and its derivative. For each i = 1, ... k we use the ith components of $\underline{Y}_{1,n}, \cdots, \underline{Y}_{n,n}, \underline{Y}_{n+1,n} = \hat{y} = \underline{x} - \underline{\lambda}_{n+1,n}$ to form:

(i) empirical estimators

$$\hat{f}^{(i)}_n(x^{(i)}) = f^{(i)}_{n,n}(y^{(i)}) = \frac{F^{(i)}_{n,n}(y^{(i)}+c_n) - F^{(i)}_{n,n}(y^{(i)}-c_n)}{2c_n}$$

$$\hat{f}'^{(i)}_n(x^{(i)}) = f'^{(i)}_{n,n}(y^{(i)}) = \frac{F^{(i)}_{n,n}(y^{(i)}+2c_n) - F^{(i)}_{n,n}(y^{(i)} - 2c_n)}{4c_n^2}$$

where $c_n \downarrow 0$, $nc_n^3 \to \infty$ and

$$F^{(i)}_{n,n}(y^{(i)}) = \frac{\#\text{ of } Y^{(i)}_{i,n}, \cdots, Y^{(i)}_{n,n} \leqslant \hat{y}^{(i)}}{n} \quad . \tag{4.37}$$

(ii) kernel estimators

$$\hat{f}^{(i)}_n(x^{(i)}) = f^{(i)}_{n,n}(y^{(i)}) = \frac{1}{na_n} \sum_{j=1}^{n} K_1^{(i)}\left(\frac{Y^{(i)}_{j,n} - \hat{y}^{(i)}}{a_n}\right) \tag{4.38}$$

$$\hat{f}'^{(i)}_n(x^{(i)}) = f'^{(i)}_{n,n}(y^{(i)}) = \frac{1}{na_n} \sum_{j=1}^{n}\left[\frac{1}{2a_n} K_2^{(i)}\left(\frac{Y^{(i)}_{j,n} - \hat{y}^{(i)}}{2a_n}\right) - \frac{1}{a_n} K_2^{(i)}\left(\frac{Y^{(i)}_{j,n} - \hat{y}^{(i)}}{a_n}\right)\right] \tag{4.39}$$

where $K_1^{(i)}$, $K_2^{(i)}$ satisfy (2.26) and (2.28) respectively.

Theorem 4.4

Suppose that for each i = 1, ... , k, $\theta^{(i)}$ is a location parameter of the density $f(x^{(i)}|\theta^{(i)})$ which is in the family (4.33). Assume that we have the observations $\underline{X}_1$, ... , $\underline{X}_n$ where for each i = 1, ... k, j = 1, ... , n, the ith component $X_j^{(i)}$ of $\underline{X}_j$ is drawn from $f(x_j^{(i)}|\theta_j^{(i)})$ and $\theta_j^{(i)}$ has the a priori distribution $G_j^{(i)}(\theta_j^{(i)})$. Define for each n = 1, 2, ... , $G_n(\underline{\theta}) = \prod_{i=1}^{k} G_n^{(i)}(\theta^{(i)})$ satisfying (4.22) and (4.25). Upon observing $\underline{X}_{n+1} = \underline{x}$ define the empirical Bayes procedure

$\hat{t}_n(\underline{x}) = a_j$ where j is an integer 1, ... , k such that

$$v^{(j)}(x^{(j)})\hat{f}_n^{(j)}(x^{(j)}) - \hat{f}_n'^{(j)}(x^{(j)}) = \max_{1\leq i\leq k}\left\{v^{(i)}(x^{(i)})\hat{f}_n^{(i)}(x^{(i)}) - \hat{f}_n'^{(i)}(x^{(i)})\right\} \tag{4.40}$$

where $v^{(i)}(x^{(i)}) = h'^{(i)}(x^{(i)})/h^{(i)}(x^{(i)})$ is known, and $\hat{f}_n^{(i)}(x^{(i)})$, $\hat{f}_n'^{(i)}(x^{(i)})$ are given by (4.36), (4.37) or (4.38), (4.39) respectively. For the problem of selecting the largest population parameter under the linear loss (4.2) we have

$$\hat{t}_n(\underline{x}) - t_{G_{n+1}}(\underline{x}) \xrightarrow{p} 0 \tag{4.41}$$

where $t_{G_{n+1}}(\underline{x})$ is given by (4.34) and furthermore $T = \{\hat{t}_n\}$ is asymptotically optimal in the sense of (3.5).

Proof

Theorems 3.4 - 3.6 guarantee (4.41). Define

$$\Delta_n(a_i,\underline{x})$$

$$= [\{v^{(1)}(x^{(1)})\hat{f}_n^{(1)}(x^{(1)}) - \hat{f}_n'^{(1)}(x^{(1)})\}-\{v^{(i)}(x^{(i)})\hat{f}_n^{(i)}(x^{(i)})$$

$$- f_n'^{(i)}(x^{(i)})\}] \cdot \prod_{i=1}^{k} \hat{f}_n^{(i)}(x^{(i)}) \quad . \tag{4.42}$$

Then we may write

$$\hat{t}_n(\underline{x}) = a_j \text{ where } j \text{ is any integer } 1, \ldots, k \text{ such that}$$

$$\Delta_n(a_j,\underline{x}) = \min_{1\leqslant i\leqslant k} \Delta_n(a_i,\underline{x}) \quad .$$

Furthermore, since

$$\prod_{i=1}^{k} \hat{f}_n^{(i)}(x^{(i)}) - \prod_{i=1}^{k} f_{G_{n+1}}^{(i)}(x^{(i)}) \xrightarrow{p} 0$$

provided $\hat{f}_n^{(i)}(x^{(i)}) - f_{G_{n+1}}^{(i)}(x^{(i)}) \xrightarrow{p} 0$ for each $i = 1, \ldots, k$ then, using Theorems 3.4 - 3.6 again,

$$\Delta_n(a_i,\underline{x}) - \Delta_{G_{n+1}}(a_i,\underline{x}) \xrightarrow{p} 0$$

where $\Delta_{G_{n+1}}(a_i,\underline{x})$ is given by (4.20), for $i = 1, \ldots k$. Hence by Theorem 4.3 $T = \{\hat{t}_n\}$ is asymptotically optimal in the sense of (3.5).

(2) Parametric Case

The method of the previous section examines the selection problem by estimating the 'present' Bayes procedure $t_{G_{n+1}}(x)$. An alternative method suggested by Robbins (1964) and considered in Chapter III, part III(5) of this thesis is to estimate the a priori distribution $G_{n+1}(\theta)$ directly. After the method of Robbins (1964) we have the following theorem. Note that for this optimality result in its general form it is not necessary to have location parameter densities for either $f(\underline{x}|\underline{\theta})$ or $G_{n+1}(\underline{\theta})$.

Theorem 4.5

Suppose we have the observations $\underline{X}_1, \ldots, \underline{X}_n$ where, for each $i=1, \ldots, k$, $j = 1, \ldots, n$, $X_j^{(i)}$ has the density $f(x_j^{(i)}|\theta_j^{(i)})$ and $\theta_j^{(i)}$ is drawn according to $G_j^{(i)}(\theta^{(i)})$. Suppose $G_n(\underline{\theta}) = \prod_{i=1}^{k} G_n^{(i)}(\theta^{(i)})$ satisfies (4.22), for all n, and that

$$(\theta^{(1)} - \theta^{(i)}) \prod_{i=1}^{k} f(x^{(i)}|\theta^{(i)}) \tag{4.43}$$

is bounded and continuous in $\underline{\theta} = (\theta^{(1)}, \ldots, \theta^{(k)})$ for all $\underline{x}$. Let $G_{n+1,n}(\underline{\theta}) = G_{n+1,n}(\underline{x}_1, \ldots, \underline{x}_n; \underline{\theta})$ be an estimate of $G_{n+1}(\underline{\theta})$ satisfying, for $i = 1, \ldots, k$,

$$P(\lim_{n\to\infty}\{G_{n+1,n}^{(i)}(\theta^{(i)}) - G_{n+1}^{(i)}(\theta^{(i)})\} = 0 \text{ at all continuity points } \theta^{(i)}) = 1 . \tag{4.44}$$

Define, for each $i=1, \ldots k$,

$$\Delta_n(a_i,\underline{x}) = \int_{\Omega^k} (\theta^{(1)} - \theta^{(i)}) f(\underline{x}|\underline{\theta}) dG_{n+1,n}(\underline{\theta}) \quad . \tag{4.45}$$

For the problem of selecting the largest population parameter under the linear loss (4.2) the procedure $T = \{t_n\}$, where

$t_n(\underline{x}) = a_j$ where j is any integer $1, \ldots k$ such that

$$\Delta_n(a_j,\underline{x}) = \min_{1\leqslant i\leqslant k} \Delta_n(a_i,\underline{x}) \quad , \tag{4.46}$$

is asymptotically optimal in the sense of (3.5).

Proof

Since by (4.20)

$$\Delta_{G_{n+1}}(a_i,\underline{x}) = \int_{\Omega^k} (\theta^{(1)} - \theta^{(i)}) f(\underline{x}|\underline{\theta}) dG_{n+1}(\underline{\theta})$$

and (4.44) implies

$$G_{n+1,n}(\underline{\theta}) - G_{n+1}(\underline{\theta}) = \prod_{i=1}^{k} G^{(i)}_{n+1,n}(\theta^{(i)}) - \prod_{i=1}^{n} G^{(i)}_{n+1}(\theta^{(i)})$$

converges, with probability one to zero at continuity points $\underline{\theta}$ of G_{n+1}, we have, upon application of the Helly-Bray theorem, that

$$\Delta_n(a_i,\underline{x}) - \Delta_{G_{n+1}}(a_i,\underline{x}) \xrightarrow{p} 0 \qquad \text{a.e.}[\underline{x}] \quad .$$

Thus, by Theorem 4.3 we have the result.

Note that if in fact $\theta^{(i)}$ is a location parameter for $f(x^{(i)}|\theta^{(i)})$ and if $G_{n+1}(\underline{\theta})$ satisfies (4.25), we have

$$G_{n+1}^{(i)}(\theta^{(i)}) = G^{*(i)}(\theta^{(i)} - \lambda_{n+1}^{(i)}) \ .$$

Thus if $G_n^{*(i)}(\theta)$ is an estimator for $G^{*(i)}(\theta)$, based on realisations of the random variables $Y_{1,n}^{(i)}, \ \ldots \ , Y_{n,n}^{(i)} = X_1^{(i)} - \lambda_{1,n}^{(i)}, \ \ldots \ , X_n^{(i)} - \lambda_{n,n}^{(i)}$, satisfying (4.44) then, as in (3.57), we use the estimate $\lambda_{n+1,n}^{(i)}$ given by (4.31), for $\lambda_{n+1}^{(i)}$, to form

$$G_{n+1,n}^{(i)}(\theta^{(i)}) = G_n^{*(i)}(\theta^{(i)} - \lambda_{n+1,n}^{(i)}) \ . \tag{4.47}$$

Provided $G^{*(i)}$ is continuous, (3.58) guarantees (4.44) for each $i = 1, \ \ldots \ , k$.

To illustrate this, we suppose that $G_{n+1}^{(i)}(\theta)$ is a member of a family characterised by the parameter $\lambda_{n+1}^{(i)}$. Estimation of $G_{n+1}(\underline{\theta})$ will therefore require only estimation of the parameter $\underline{\lambda}_{n+1} = (\lambda_{n+1}^{(1)}, \ \ldots \ , \lambda_{n+1}^{(k)})$. In particular suppose $f(x_j^{(i)} \mid \theta_j^{(i)})$ has a $N(\theta_j^{(i)}, \beta^{2(i)})$ density, where $\theta_j^{(i)}$ is drawn from the a priori distribution $G_j^{(i)}(\theta^{(i)})$ which has a $N(\lambda_j^{(i)}, \sigma^{2(i)})$ density, for $i = 1, \ \ldots \ , k$, $j = 1, \ \ldots \ , n+1$. We again assume that $\lambda_j^{(i)}$ satisfies (4.31). In order to make use of Theorem 4.5 we note that

$$(\theta^{(1)} - \theta^{(i)}) \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi}\ \sigma^{(i)}} \exp\left\{-\tfrac{1}{2}\left[\frac{x^{(i)} - \theta^{(i)}}{\sigma^{(i)}}\right]^2\right\}$$

is clearly bounded and therefore (4.43) is satisfied. Also, if $G^{*(i)}(\cdot)$ is a distribution function with a $N(0, \sigma^{2(i)})$ density then we have that

$$G_{n+1}^{(i)}(\theta^{(i)}) = G^{*(i)}(\theta^{(i)} - \lambda_{n+1}^{(i)})$$

is the 'present' a priori distribution. In particular

$$g_{n+1}^{(i)}(\theta^{(i)}) = \frac{1}{\sqrt{2\pi}\,\sigma^{(i)}} \exp\left\{-\tfrac{1}{2}\left[\frac{\theta^{(i)} - \lambda_{n+1}^{(i)}}{\sigma^{(i)}}\right]^2\right\} .$$

Thus replacing $\lambda_{n+1}^{(i)}$ by $\lambda_{n+1,n+1}^{(i)}$, defined by (4.31), for $i = 1, \ldots, k$ we have the estimator

$$G_{n+1,n+1}^{(i)}(\theta^{(i)}) = G_{n+1}^{*(i)}(\theta^{(i)} - \lambda_{n+1,\,n+1}^{(i)}) .$$

This guarantees (4.44) and Theorem 4.4 thus gives the asymptotically optimal procedure for the selection problem under the linear loss (4.2) as (4.46), which for this Normal-Normal example reduces to

$t_n(\underline{x}) = a_j$ where j is any integer $1, \ldots, k$ such that

$$\frac{\sigma^{2(j)}x^{(j)} + \beta^{2(j)}\lambda_{n+1,n+1}^{(j)}}{\sigma^{2(j)} + \beta^{2(j)}} = \max_{1\leqslant i\leqslant k}\left\{\frac{\sigma^{2(i)}x^{(i)} + \beta^{2(i)}\lambda_{n+1,n+1}^{(i)}}{\sigma^{2(i)} + \beta^{2(i)}}\right\} . \tag{4.48}$$

Note also that this procedure is asymptotically optimal for the selection problem under zero-one loss. Further parametric examples of this type have been given by Deely (1965) and these could be adapted to the present changing prior case noting again, that with such strong parametric assumptions on the priors $\{G_n\}$, it is not necessary to restrict $f(x^{(i)}|\theta^{(i)})$ and $g_n^{(i)}(\theta^{(i)})$ to being location parameter densities.

CHAPTER V

I. APPLICATION TO A RELIABILITY MODEL

Suppose we have a system made up of several components which operates in continuous time. The general study of the reliability of such systems is a well established problem. Briefly, the measure of reliability for a system is the frequency at which failures of the system may occur. In particular, the problem of reliability is to express numerically the probabilities that the system will operate for specified lengths of time in the environment for which it is designed. In general, reliability distinguishes three characteristic types of failures:

(i) Failures which occur early in the life of a component. These can be eliminated by a 'burn-in' testing of the system.

(ii) Failures caused by wearout of the components. Reliability against this type of failure can be achieved by replacing those components at regular intervals and to make the replacement interval shorter than the mean wear out life of the component.

(iii) Random failures. The occurrence of these failures cannot be predicted but it is assumed that the frequency of their occurrence during sufficiently long periods can be probabilistically determined.

Suppose now that the time between successive failures in the system is a random variable, X, distributed conditionally according to the density $f_\theta(x)$. Characteristically, θ is some kind of measure of the failure rate of the system. A very convenient measure of the reliability of the system is the mean time between failures, namely $E[X|\theta]$ and a failure rate is therefore its reciprocal. Another function often considered is the hazard rate

$$f_\theta(x)/(1-F_\theta(x))$$

which is just a failure rate measure conditional upon past survival.

Most reliability investigators are interested in obtaining some kind of point estimate or confidence bounds for the parameter θ when the functional form of the failure density has been specified and for which some kind of life-test data are available. The Bayesian approach is thus to assume that the parameter θ has a known a priori distribution $G(\theta)$. This approach to reliability analysis has received attention from many authors and a survey of these results may be found in Shimi and Tsokos (1977).

The loss function we use here is squared error loss. Thus on the basis of the observation X=x, we have the estimator

$$t_G(x) = E_G[\theta|X=x] \quad .$$

The empirical Bayes approach is then to assume that $G(\theta)$ is unknown, but that prior history is available in the form of

the life data $X_1, \ldots, X_n$. Estimates of $G(\theta)$ and $t_G(x)$ are then made on the basis of these observations according to the procedures described in Chapters I and II.

The above model assumes that the process is basically a stable one. That is, the failure measure θ is determined by a constant prior G throughout the history of the system. Suppose however that due to improvements in the process which manufactures the components for our system, the reliability of the system improves. Alternatively, if the system fails and is subsequently repaired, it is reasonable to assume that the failure rate of our system has been affected in some way. That is we assume that the failure rate measure θ exhibits a monotonic change. This may be expressed as a change in the quantity $P(\theta < a)$ for all $a \in \Omega$, i.e. G changes.

Formally, suppose $X_1, \ldots, X_n$ are independent random variables, where X_i denotes the working time of the system between the repair of failure i -1 and the occurrence of failure i. Each X_i is distributed according to $f(x_i|\theta_i)$ where θ_i is a failure rate measure for the system between failures i-1 and i, $i = 1, \ldots, n$. The a priori distribution of θ_i is $G_i(\theta)$, $i = 1, \ldots, n$.

(1) A Wear-Out Model

Suppose we have a system subject to wear-out failures. When a component fails it is immediately replaced. The reliability of the system is thereby affected. In fact the change in reliability may be a monotone one due to an increase in the probability of failure with the age of the system.

Assume that the wear-out time is governed by the density

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\theta}{\sigma}\right)^2} \quad . \tag{5.1}$$

This density usually approximates wear-out phenomena quite well with half of the failures occurring before the mean time between failures, θ, with a clustering around θ.

Thus, if we further assume that $G_i(\theta)$ has a location parameter $\lambda_i = E_{G_i}[\theta]$, for all i, then at stage n+1 we observe the time to the (n+1)st failure $X_{n+1} = X$, and the Bayes procedure for estimating the mean time θ_{n+1} is

$$t_{G_{n+1}}(x) = x + \sigma^2 \left(\frac{f^{(1)}_{G_{n+1}}(x)}{f_{G_{n+1}}(x)} \right) \quad . \tag{5.2}$$

As before define $Y_i = X_i - \lambda_i$. This gives the results of Chapter 3, namely

$$G_i(\theta) = G^*(\theta - \lambda_i)$$

$$f_{G_i}(x) = f_{G^*}(y) \quad .$$

Consequently with $y = x - \lambda_{n+1}$, the Bayes procedure for this problem becomes

$$t_{G^*}(y) = (y + \lambda_{n+1}) + \sigma^2 \frac{f^{(1)}_{G^*}(y)}{f_{G^*}(y)} \quad . \tag{5.3}$$

Thus, provided $G_{n+1}(\theta)$ is known we have an estimator for the mean time between failures n and n+1.

The empirical Bayesian now postulates that $G_{n+1}(\theta)$ is not known but that prior history in the form of the lifetime

observations $X_1, \ldots, X_n$ is available. Again assume that the change in the mean of the a priori distributions is of the linear form (3.15). Since the $\{\lambda_i\}$ sequence is unknown we form the estimates $\lambda_{i,n}$ given by (3.19).

Since estimation of the Bayes procedure now involves only estimation of $f_{G*}(y)$ and $f^{(1)}_{G*}(y)$, we use the estimators $f_{n,n}(y)$ and $f^{(1)}_{n,n}(y)$ of (3.30) and (3.31) or (3.36) and (3.39) respectively to give

$$\hat{t}_n(x) = t_n(y) = (y + \lambda_{n+1}) + \sigma^2 \frac{f^{(1)}_{n,n}(y)}{f_{n,n}(y)} \quad . \tag{5.4}$$

If we now define $s_n(x)$ by (3.49) it follows from Theorem 3.9 that $T = \{s_n\}$ is asymptotically optimal in the sense of (3.5). Alternatively using the kernel estimators (3.36) and (3.39) we define the estimator $\bar{t}_n(y)$ by (3.72) and a rate of convergence result is evident. Furthermore if, as in section IV(2) of Chapter III, we assume that $G_n(\theta)$ is a parametric distribution, then convergence rates will be more readily forthcoming.

So far no assumption has been made on the way on which $\{z_n\}$ (and therefore $\{\lambda_n\}$) changes. Suppose $\{z_n\}$ is a monotone decreasing sequence. Then for the location parameter distributions $G_i(\theta)$ we have

$$P(\theta_i \leqslant u) = P(\theta_i - \lambda_i \leqslant u - \lambda_i)$$

$$= P(\theta_{i+1} - \lambda_{i+1} \leqslant u - \lambda_i) \quad , \quad \text{since}$$

the $\{\theta_i - \lambda_i\}$ are i.i.d. random variables

$$= P(\theta_{i+1} \leqslant t + \lambda_{i+1} - \lambda_i)$$

$$\leqslant P(\theta_{i+1} \leqslant t) \quad , \quad \text{since } \lambda_{i+1} - \lambda_i$$

$= a(z_{i+1} - z_i) < 0$, for $a > 0$. That is

$$G_{i+1}(u) \geqslant G_i(u) \quad . \tag{5.5}$$

This means that $P(\theta_i > u)$ is monotonically decreasing with i, so that we have a case of reliability degradation with the mean time between failure θ_i stochastically decreasing.

(2) A Random Failure Model

Consider a system which is subject only to failures which occur at random intervals. The life of the system between failures i-1 and i is a random variable X_i distributed according to the exponential density

$$f_{\theta_i}(x) = \begin{cases} \theta_i e^{-\theta_i x} & , \quad x > 0, \quad \theta_i > 0 \\ 0 & , \quad \text{elsewhere} \quad . \end{cases} \tag{5.6}$$

The parameter θ_i is here the failure rate of the system, and the mean time between failures is thus $1/\theta_i$. We suppose as before that the failure rate changes stochastically. In particular suppose that θ_i has the a priori distribution $G_i(\theta)$ where

$$G_i'(\theta) = g_i(\theta) = \begin{cases} \dfrac{\beta_i^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta_i \theta} & , \quad \theta > 0 \\ 0 & , \text{ elsewhere} \quad . \end{cases} \tag{5.7}$$

Note that since β_i is a scale parameter of this distribution

$$\beta_i \leqslant \beta_{i+1} \Rightarrow G_i(\theta) \leqslant G_{i+1}(\theta), \qquad \forall\ \theta \in \Omega \quad .$$

That is monotonicity of $\{\beta_i\}$ guarantees the stochastic monotonicity of the failure rates $\{\theta_i\}$. In fact if $\{\beta_i\}$ decreases, $\{\theta_i\}$ becomes stochastically smaller thus corresponding to a situation of reliability degradation.

For this model, if complete knowledge is assumed concerning the priors $\{G_i\}$ then the Bayesian estimate for the failure rate of the system at stage n+1 (i.e. between the nth and (n+1)st failures) under squared error loss is

$$t_{G_{n+1}}(x) = \frac{\alpha+1}{\beta_{n+1}+x} \tag{5.8}$$

Suppose however that $\{\beta_i\}$ is not known but that its structure is known to be of the linear form

$$\beta_i = az_i \tag{5.9}$$

where $\{z_i\}$ is a known sequence. It is also assumed that past evidence is available in the form of the observed lifetimes $X_1, \ldots, X_n$. Thus in order to determine system reliability we need to make inference about a.

Note that for the density (5.6) and for all i,

$$E[X_i|\theta_i] = \frac{1}{\theta_i}, \quad \text{giving}$$

$$E[X_i] = E_{G_i}[E[X_i|\theta_i]] = E_{G_i}\left|\frac{1}{\theta_i}\right| = \frac{\beta_i}{\alpha-1} \quad . \tag{5.10}$$

A simple estimator for the parameter a, based on $X_1, \ldots, X_n$, is

$$a_n = \frac{1}{n}\sum_{i=1}^{n}\frac{(\alpha-1)X_i}{z_i} \quad . \tag{5.11}$$

Lemma 5.1

With $\beta_{n+1,n} = a_n z_{n+1}$ we have

$$E|\beta_{n+1,n} - \beta_{n+1}|^2 = \beta_{n+1}^2/(\alpha-2)n \tag{5.12}$$

Proof

$$E|\beta_{n+1,n} - \beta_{n+1}|^2 = Z_{n+1}^2 E|a_n - a|^2$$

$$= z_{n+1}^2 E\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{(\alpha-1)X_i}{z_i} - \frac{(\alpha-1)E[X_i]}{z_i}\right)\right|^2$$

$$= \frac{z_{n+1}^2(\alpha-1)^2}{n^2}\sum_{i=1}^{n} E\left|\frac{X_i - E(X_i)}{z_i}\right|^2 \quad . \tag{5.13}$$

Now Var $(X_i) = \beta_i^2/(\alpha-2)$ so that (5.13) becomes

$$= \frac{z_{n+1}(\alpha-1)^2}{n^2} \cdot \frac{1}{(\alpha-1)^2(\alpha-2)} \cdot \sum_{i=1}^{n}\frac{\beta_i^2}{z_i^2}$$

$$= \frac{z_{n+1}^2 a^2}{(\alpha-2).n} = \frac{\beta_{n+1}^2}{(\alpha-2).n}$$

as required.

Since $\beta_i > 0$ for all i, let $B = \sup_i \left\{\frac{1}{\beta_i}\right\} < \infty$.

Then define

$$\hat{\beta}_{n+1,n} = \max\{\beta_{n+1,n}, B\} \quad .$$

This gives the the empirical Bayesian estimator for the failure rate θ_{n+1} as

$$t_n(x) = \frac{\alpha+1}{\hat{\beta}_{n+1,n} + x} \quad . \tag{5.14}$$

The following lemma gives asymptotic optimality for the procedure (5.14) upon application of Lemma 1.3.

Lemma 5.2

$$E|t_n(x) - t_{G_{n+1}}(x)|^2 \leqslant \frac{(\alpha+1)^2}{(\alpha-2)} B \cdot \frac{1}{n} \tag{5.15}$$

Proof

$$E|t_n(x) - t_{G_{n+1}}(x)|^2 = (\alpha+1)^2 E\left| \frac{\hat{\beta}_{n+1,n} - \beta_{n+1}}{(\hat{\beta}_{n+1,n}+x)(\beta_{n+1}+x)} \right|^2$$

$$\leqslant (\alpha+1)^2 E \left| \frac{\beta_{n+1,n} - \beta_{n+1}}{(\hat{\beta}_{n+1,n})\beta_{n+1}} \right|^2$$

$$\leqslant \frac{(\alpha+1)^2}{\beta_{n+1}^2} E|\beta_{n+1,n} - \beta_{n+1}|^2 \cdot E\left|\frac{1}{\hat{\beta}_{n+1}^2}\right|$$

$$\leqslant \frac{(\alpha+1)^2 \cdot \beta_{n+1}^2}{\beta_{n+1}^2 (\alpha-2) \cdot n} \cdot B, \text{ by Lemma 5.1.}$$

Hence the result.

In the models examined above we are interested in determining the failure rate after an actual failure had occurred. However, often future planning of the maintenance of the system is required. For example in reliability growth situations, it may be necessary to have some idea of how long it will be before a specified reliability is achieved. In reliability decay, the same problem is concerned with the length of useful life of a system before it becomes so unreliable that it must be withdrawn from service. In this case it becomes necessary to make some prediction for the failure rate at some future time. This may involve predicting either the failure rate θ or the distribution of the next time to failure. Littlewood and Verrall (1974) consider a monotone Bayesian reliability model in which they are interested in determining the future behaviour of the system. They examine the unconditional distribution of the time between the nth and the (n+1)st failures, X_{n+1}. For the exponential-gamma case this is

$$f_{G_{n+1}}(x) = \int_{\Omega} f(x|\theta)\, dG_{n+1}(\theta)$$

$$= \frac{\alpha \beta_{n+1}^{\alpha}}{(\beta_{n+1} + x)^{\alpha+1}} \quad , \quad x > 0 \quad . \qquad (5.16)$$

This p.d.f. may then be used to determine a measure of reliability such as the instantaneous failure rate (or hazard rate) of the system.

In the Bayesian case if α and β_{n+1} are both known then so too is $f_{G_{n+1}}(x)$. However if β_{n+1} is unknown, but it is

assumed that it changes linearly with n according to (5.9), then we use the estimator $\beta_{n+1,n}$ based on $X_1, \ldots, X_n$ given by $\beta_{n+1,n} = a_n z_{n+1}$ where a_n is given by (5.11). Since $\{z_n\}$ is a known sequence this method may be used for the prediction of any future value β_{n+k} by using $\beta_{n+k,n} = a_n z_{n+k}$. Convergence of these estimates guarantees convergence of the respective densities (5.16).

CHAPTER VI

EXTENSION TO THE CASE OF A STOCHASTICALLY CHANGING PRIOR DISTRIBUTION

I. INTRODUCTION

The main content of this thesis deals with the case in which the change in the a priori distribution is assumed to follow a given deterministic pattern - in particular a linear drift in the mean. Suppose now that the means change in some probabilistic fashion. For example consider again the problem of determining the fraction defective θ in a consignment of resistors, as mentioned in Chapter III. This model in which the prior changes stochastically would now be appropriate to a situation in which the manufacturing process receives random shocks.

Thus consider the usual sequence $(\theta_1, X_1), \ldots, (\theta_n, X_n)$ where the X_i are independent and, given $\theta_i = \theta$, drawn according to the density $f(x|\theta)$. For each i, θ_i has the a priori distribution $G_i(\theta)$, each G_i having the same support, Ω. Suppose $\{\Lambda_i\}_{i=1}^{\infty}$ is a sequence of random variables defined on the probability space $(S', \mathcal{F}', Q)$. Based on the realisations $\Lambda_i = \lambda_i$, i = 1,2, ... we consider the case where $G_i(\theta)$ is a member of a parametric family indexed by the parameter λ_i.

Consider the problem for which (3.10) and (3.11) hold, with

$$\int_{\Omega} \theta dG_i(\theta) = \lambda_i \quad . \tag{6.1}$$

If we know the values $\lambda_1,\ldots,\lambda_n$, λ_{n+1} then we know exactly what the a priori distribution is at each stage for this parametric case. That is we are able to take the Bayes procedure at each stage and the problem is solved.

Suppose now that we do not know $\{\lambda_i\}$. At stage n+1 we have the observations $X_1 = x_1, \ldots, X_n = x_n$, and we observe $X_{n+1} = x$. We will estimate the 'present' a priori distribution $G_{n+1}(\theta) = G_{n+1}(\theta;\lambda_{n+1})$ by replacing λ_{n+1} with an estimate $\lambda_{n+1,n+1}$ to give

$$G_{n+1,n+1}(\theta) = G_{n+1}(\theta;\lambda_{n+1,n+1}) \quad . \tag{6.2}$$

This has the advantage that the a priori estimator is also in the same parametric family as G_{n+1}. The empirical Bayes procedure is now

$$t_n(x) = t_n(x_1, \ldots, x_n;x)$$

$$= t_{G_{n+1,n+1}}(x) \quad , \tag{6.3}$$

and it remains to verify (3.7) to give asymptotic optimality by Theorems 3.1 and 3.2. Note also that as mentioned earlier, when a parametric assumption is made about $\{G_i\}$, it is not necessary to assume the conditions (3.10), (3.11).

Since λ_i is the realisation of a random variable Λ_i, any probability statements made about the estimators $\lambda_{n+1,n+1}$ must be made with reference to the measure Q. Conditional upon $\Lambda_i = \lambda_i$, X_i is a random variable defined on the probability

space $(S, \mathcal{F}, \mu)$ by assumption. X_i unconditionally may thus be considered as a random variable defined on the product space $(S \times S', \mathcal{F} \times \mathcal{F}', \mu \times Q)$. Therefore any estimate $\lambda_{n+1,n+1}$ of λ_{n+1} is a realisation of a random variable $\Lambda_{n+1,n+1}$ defined on this product space. We thus consider the two forms of asymptotic optimality.

Definition 1

Let $T = \{t_n\}$ where $t_n(x)$ is given by (6.3). Suppose R_n, R are as given in Chapter 1 and if

$$\lim_{n\to\infty} \{R_n(T,G_{n+1}) - R(G_{n+1})\} = 0 \qquad \text{a.e.}[Q] \qquad (6.4)$$

then we say that T is asymptotically optimal.

An alternative definition, involving an overall expectation, is

Definition 2

Let $T = \{t_n\}$ be given by (6.3). If

$$\lim_{n\to\infty} E_Q[R_n(T,G_{n+1}) - R(G_{n+1})] = 0 \qquad (6.5)$$

where E_Q denotes expectation relative to the probability measure Q, then T is said to be asymptotically optimal.

II OPTIMALITY RESULTS

If there is no relationship between sucessive Λ_i's then the problem is a pure Bayesian one - the previous

observations are not useful for determining the 'present' optimal procedure. Suppose however that Λ_i is determined stochastically as follows:

$$\Lambda_i = \Lambda_{i-1} + \delta_i$$

where δ_i is a random variable defined on $(S', \mathcal{F}', Q)$ having zero mean and variance η_i^2. That is if $\mathcal{F}_i$ is the σ-algebra generated by $\Lambda_1, \ldots, \Lambda_i$, $\mathcal{F}_i \subset \mathcal{F}'$ and $\{(\Lambda_i, \mathcal{F}_i)\}$ forms a martingale sequence. We assume that

$$\sup_i E_Q|\Lambda_i| < \infty \quad . \tag{6.6}$$

We require an estimator for the 'present' parameter random variable Λ_{n+1} using $X_1, \ldots, X_n, X_{n+1}$. Suppose

$$\mathrm{Var}(X_i|\Lambda_i) = S_i^2 \qquad , \qquad \text{for each } i,$$

which may or may not depend upon Λ_i.

Theorem 6.1

Suppose that (6.1) holds and that $X_1, \ldots, X_n, X_{n+1}$ are drawn from $f(x|\theta)$ for which

$$E(X|\theta) = \theta \quad . \tag{6.7}$$

For each $s' \in S'$ with $\Lambda_i(s') = \lambda_i$, $S_i^2(s') = \sigma_i^2$, $i = 1,2, \ldots$, suppose $\mathrm{Var}(X_i|\lambda_i) = \sigma_i^2$ is such that

$$\sum_{i=1}^{\infty} \sigma_i^2/i^2 < \infty \quad . \tag{6.8}$$

Define the random variable

$$\Lambda_{n+1,n+1} = \frac{1}{n+1} \sum_{i=1}^{n+1} X_i \quad . \tag{6.9}$$

We have

$$E|\Lambda_{n+1,n+1} - \Lambda_{n+1}|^2 \to 0 \qquad \text{a.e.[Q]},$$

where E denotes expectation over $X_1, \ldots, X_{n+1}$ with respect to μ.

Proof

The submartingale convergence theorem (see Ash (1972)) provides the existence of a random variable Λ defined on $(S', \mathcal{F}', Q)$ such that

$$\Lambda_n \to \Lambda \qquad \text{a.e.[Q]} \quad . \tag{6.10}$$

Note that

$$\Lambda_n = \sum_{i=1}^{n} (\Lambda_i - \Lambda_{i-1}) + \Lambda_o \quad .$$

From Apostol (1974) convergence of the series Λ_n implies Césaro convergence i.e.

$$\frac{1}{n} \sum_{i=1}^{n} \Lambda_i \to \Lambda \qquad \text{a.e.[Q]} \quad . \tag{6.11}$$

For each $s' \in S'$, with $\Lambda_i(s') = \lambda_i$, $i = 1,2, \ldots$, the $X_i - \lambda_i$ are independent random variables with finite variances $S_i^2(s') = \sigma_i^2$, satisfying (6.8), and zero mean. Thus by the Strong Law of Large Numbers, (Lukacs (1975), Theorem 3.4.1),

$$\frac{1}{n} \sum_{i=1}^{n} (X_i - \lambda_i) \to 0 \qquad \text{a.e.}[\mu] \quad . \tag{6.12}$$

Thus (6.11), (6.12) combine to give

$$\frac{1}{n} \sum_{i=1}^{n} X_i \to \Lambda \qquad \text{a.e.}[\mu \times Q] . \tag{6.13}$$

Furthermore we have

$$E|\Lambda_{n+1,n+1} - \Lambda_{n+1}|^2 = E\left|\frac{1}{n+1} \sum_{i=1}^{n+1} X_i - \Lambda_{n+1}\right|^2$$

and since $E[X_i - \Lambda_i] = 0$,

$$= E\left|\frac{1}{n+1} \sum_{i=1}^{n+1} (X_i - \Lambda_i)\right|^2 + \left|\frac{1}{n+1} \sum_{i=1}^{n+1} \Lambda_i - \Lambda_{n+1}\right|^2$$

$$= \frac{1}{(n+1)^2} \sum_{i=1}^{n+1} S_i^2 + \left|\frac{1}{n+1} \sum_{i=1}^{n+1} \Lambda_i - \Lambda_{n+1}\right|^2 \quad . \tag{6.14}$$

Under assumption (6.8), the first term of (6.14) converges to zero a.e.[Q] upon application of the Kronecker Lemma, (see Stout (1974)). Using (6.10) and (6.11), the second term of (6.14) converges to zero a.e.[Q] to give the result.

Remark

This theorem is used to give asymptotically optimal procedures $t_{G_{n+1,n+1}}(x)$ for the squared error loss estimation

problem in the sense that

$$E\left|t_{G_{n+1,n+1}}(X) - t_{G_{n+1}}(X)\right|^2 \to 0 \qquad \text{a.e.}[Q] .$$

For more general loss structures under the restriction

$$\int_{\Omega} L(\theta)\, dG_n(\theta) < \infty \qquad \text{a.e.}[Q]$$

where $L(\theta) = \sup_{a \in A} L(a,\theta)$, all that is required is

$$\Lambda_{n+1,n+1} - \Lambda_{n+1} \to 0 \qquad \text{a.e.}[Q]$$

with probability one with respect to μ. This guarantees that $G_{n+1,n+1}(\theta) - G_{n+1} \to 0$ a.e.[Q] with probability one, for all θ. In fact

$$\Lambda_{n+1,n+1} - \Lambda_{n+1} = \frac{1}{n+1} \sum_{i=1}^{n+1} X_i - \Lambda_{n+1}$$

$$= \frac{1}{n+1} \sum_{i=1}^{n+1} X_i - \Lambda + \Lambda_{n+1} - \Lambda$$

which converges with probability one a.e.[Q] by (6.10) and (6.13).

Since a.e. convergence does not guarantee convergence in square mean (i.e. in L_2), which is needed for optimality results in terms of the overall expectation, we introduce the concept of uniform integrability.

Definition

Let $h_1, h_2, \ldots$ be defined on the probability space $(S', \mathcal{F}', Q)$. The h_i are said to be uniformly integrable iff

$$\int_{\{h_i \geqslant c\}} |h_i| dQ \to 0, \text{ as } c \to \infty \quad . \tag{6.15}$$

The following theorem gives the general condition for convergence in square mean.

Theorem 6.2

Let $h_1, h_2, \ldots$ be defined on the probability space $(S', \mathcal{F}', Q)$. If $h_n \to h$ in probability and if the $|h_i|^2$ are uniformly integrable then

$$\int_{S'} |h_n - h|^2 dQ \to 0 \quad . \tag{6.16}$$

The proof of this result is found in Ash (1972, p297). To simplify this theorem for application to the present problem the following lemma is useful.

Lemma 6.3

Let Λ_i, $i = 1, 2, \ldots$, be integrable random variables on $(S', \mathcal{F}', Q)$. If for some $p > 1$

$$E_Q|\Lambda_i|^p \leqslant M < \infty, \text{ for all } i = 1, 2, \ldots, \tag{6.17}$$

the Λ_i are uniformly integrable.

Proof

Given $\varepsilon > 0$, let $M = \sup_i E_Q|\Lambda_i|^p$ and set $a = M/\varepsilon$. There exists a positive number c such that $t^p/t \geqslant a$ for $t \geqslant c$. Thus

$$\int_{\{\Lambda_i \geqslant c\}} |\Lambda_i| dQ \leqslant \int_{\{\Lambda_i \geqslant c\}} \frac{|\Lambda_i|^p}{a} dQ$$

$$\leqslant \int_{S'} \frac{|\Lambda_i|^p}{a} dQ \leqslant \frac{M}{a} = \varepsilon \quad .$$

Hence the result.

For the present problem we have:

Theorem 6.4

Suppose the conditions of Theorem 6.1 hold with the additional assumption

$$E_Q|\Lambda_i|^{2t} \leqslant M < \infty \quad , \tag{6.18}$$

for some $t > 1$. Then

$$E_Q\{E|\Lambda_{n+1,n+1} - \Lambda_{n+1}|^2\} \to 0 \quad .$$

Proof

From (6.14)

$$E_Q\{E|\Lambda_{n+1,n+1} - \Lambda_{n+1}|^2\}$$

$$= E_Q\left|\frac{1}{(n+1)^2}\sum_{i=1}^{n+1} S_i^2\right| + E_Q\left|\frac{1}{n+1}\sum_{i=1}^{n+1}\Lambda_i - \Lambda_{n+1}\right|^2 \tag{6.19}$$

$$\leqslant E_Q\left|\frac{1}{(n+1)^2}\sum_{i=1}^{n+1} S_i^2\right| + 2\{E_Q\left|\frac{1}{n+1}\sum_{i=1}^{n+1}\Lambda_i - \Lambda\right|^2$$

$$+ E_Q|\Lambda_{n+1} - \Lambda|^2\} \ . \qquad (6.20)$$

The first term of (6.20) converges to zero since

$$\frac{1}{(n+1)^2}\sum_{i=1}^{n+1} S_i^2 \to 0 \qquad \text{a.e.[Q]} \ .$$

From (6.18) and Lemma 6.3 applied to $|\Lambda_i|^2$ we see that the $|\Lambda_i|^2$ are uniformly integrable and since $\Lambda_n \to \Lambda$ a.e.[Q] we have

$$E_Q|\Lambda_{n+1} - \Lambda|^2 \to 0 \ ,$$

by Theorem 6.2. To complete the proof it is necessary to show only that the $\frac{1}{n+1}\sum_{i=1}^{n+1}|\Lambda_i|^2$ are uniformly integrable, since

$$\frac{1}{n+1}\sum_{i=1}^{n+1}\Lambda_i \to \Lambda \qquad \text{a.e.} \quad \text{That is, we need}$$

$$E_Q\left|\frac{1}{n}\sum_{i=1}^{n}\Lambda_i\right|^{2t} \leqslant M < \infty \qquad (6.21)$$

for some $t > 1$, by Lemma 6.3. In particular

$$\left\{E_Q\left|\frac{1}{n+1}\sum_{i=1}^{n+1}\Lambda_i\right|^{2t}\right\}^{\frac{1}{2t}} = \frac{1}{n+1}\left\{E_Q\left|\sum_{i=1}^{n+1}\Lambda_i\right|^{2t}\right\}^{\frac{1}{2t}}$$

$$\leqslant \frac{1}{n+1}\sum_{i=1}^{n+1}\left\{E_Q|\Lambda_i|^{2t}\right\}^{\frac{1}{2t}} \ ,$$

by repeated application of Minkowski's inequality,

$$\leqslant \frac{1}{n+1}\left((n+1)M^{\frac{1}{2t}}\right)$$

thus satisfying (6.12). Again applying Theorem 6.2 we have

$$E_Q\left|\frac{1}{n+1}\sum_{i=1}^{n+1}\Lambda_i - \Lambda\right|^2 \to 0 \quad ,$$

yielding the result.

In order to recover a rate of convergence in square mean it will be necessary to make assumptions about the order of the variances of the Λ_i, in particular assumptions on the η_i^2. In addition a strengthening of (6.8) is needed.

Theorem 6.5

Suppose the conditions of Theorem 6.1 are valid with

$$\sum_{i=1}^{\infty} i^{\delta}\eta_i^2 \to \eta^2 < \infty \tag{6.22}$$

for some $0 < \delta < 2$ where $\eta_i^2 = \text{Var}\ (\Lambda_i - \Lambda_{i-1})$. Also suppose that

$$\sum_{i=1}^{n+1} \frac{S_i^2}{i^{2-\beta}} < \infty \qquad \text{a.e.[Q]}$$

for some $\beta > 0$. Then

$$E_Q\{E|\Lambda_{n+1,n+1} - \Lambda_{n+1}|^2\} = 0((n+1)^{-\gamma}) \quad , \tag{6.23}$$

where $\gamma = \min\{\delta,\beta\}$.

Proof

Since $\Lambda_i - \Lambda_{n+1} = \sum_{j=i}^{n} (\Lambda_j - \Lambda_{j+1})$, we have

$$\sum_{i=1}^{n+1} (\Lambda_i - \Lambda_{n+1}) = \sum_{i=1}^{n} i(\Lambda_i - \Lambda_{i+1}) \quad .$$

Thus, from (6.14)

$$E_Q\{E|\ \Lambda_{n+1,n+1} - \Lambda_{n+1}|^2\}$$

$$= E_Q\left|\frac{1}{(n+1)^2} \sum_{i=1}^{n+1} S_i^2\right| + \frac{1}{(n+1)^2} E_Q\left|\sum_{i=1}^{n} i(\Lambda_i - \Lambda_{i+1})\right|^2 \quad . \tag{6.24}$$

For a martingale $\{\Lambda_i, \mathcal{F}_i\}$, the differences $\Lambda_j - \Lambda_{j+1}$ and $\Lambda_k - \Lambda_{k+1}$, $j \neq k$, are mutually orthogonal and thus the second term of (6.24) becomes

$$\frac{1}{(n+1)^2} \sum_{i=1}^{n} E_Q[i^2(\Lambda_i - \Lambda_{i+1})^2]$$

$$= \frac{1}{(n+1)^2} \sum_{i=1}^{n} i^2\eta_i^2$$

$$\leqslant \frac{1}{(n+1)^\delta} \sum_{i=1}^{n} i^\delta \eta_i^2 \leqslant \frac{\eta^2}{(n+1)^\delta} \quad .$$

Furthermore

$$\frac{1}{(n+1)^2} \sum_{i=1}^{n+1} S_i^2 \leqslant \frac{1}{(n+1)^\beta} \sum_{i=1}^{n+1} S_i^2/i^{2-\beta} = 0((n+1)^{-\beta}).$$

Hence the result follows from (6.24).

To illustrate the above results consider the following example. Assume $\{\Lambda_i, \mathcal{F}_i\}$ is a martingale satisfying (6.6).

(1) Normal-Normal

Suppose that $f(x|\theta)$ has a $N(\theta,\sigma_1^2)$ distribution function with σ_1^2 known. Given $\Lambda_i = \lambda_i$, $G_i(\theta)$ has a $N(\lambda_i,\sigma_2^2)$ distribution function. As before this yields $f_{G_i}(x|\lambda_i) \sim N(\lambda_i, \sigma^2)$, where $\sigma^2 = \sigma_1^2 + \sigma_2^2$. That is, $S_i^2 = \sigma^2 = \text{constant}$ a.e.[Q]. If the value $\Lambda_{n+1} = \lambda_{n+1}$ was known to us, then we know the a priori distribution $G_{n+1}(\theta) = G_{n+1}(\theta;\lambda_{n+1})$ exactly and thus we have the Bayes procedure for any given decision problem. However, with λ_{n+1} unknown we use the previous observations to form the random variable $\Lambda_{n+1,n+1}$ as in (6.8).

For each $s' \in S'$ with $\Lambda_{n+1,n+1}(s') = \lambda_{n+1,n+1}$ we have

$$g_{n+1,n+1}(\theta) = G'_{n+1,n+1}(\theta) = G'_{n+1}(\theta;\ \lambda_{n+1,n+1})$$

$$= \frac{1}{\sqrt{2\pi}\ \sigma_2}\ \exp\left\{-\tfrac{1}{2}\left(\frac{\theta - \lambda_{n+1,n+1}}{\sigma_2}\right)^2\right\} \quad .$$

Thus since

$$\Lambda_{n+1,n+1} - \Lambda_{n+1} \to 0 \qquad \text{a.e.}[\mu x Q]$$

we have

$$P(\lim_{n\to\infty}\{G_{n+1,n+1}(\theta) - G_{n+1}(\theta)\} = 0 \text{ at continuity points}$$

$$\theta \text{ of } G_{n+1}) = 1 \qquad \text{a.e.}[Q] \quad , \qquad (6.25)$$

where P is the probability over $X_1, \ldots, X_n, X_{n+1}$ with respect to the measure μ.

(a) The Two-Action Problem

Consider the hypothesis testing problem (3.41) under the piecewise linear loss structure (1.32). For each $s' \in S'$ with $\Lambda_{n+1}(s') = \lambda_{n+1}$ and $\Lambda_{n+1,n+1}(s') = \lambda_{n+1,n+1}$ define

$$\Delta_{G_{n+1}}(x) = \int_{-\infty}^{\infty} (\theta_o - \theta) f(x|\theta) dG_{n+1}(\theta) \quad ,$$

$$\Delta_{G_{n+1,n+1}}(x) = \int_{-\infty}^{\infty} (\theta_o - \theta) f(x|\theta) dG_{n+1,n+1}(\theta) \quad .$$

Since $(\theta_o - \theta) \frac{1}{\sqrt{2\pi}\,\sigma_1} \exp\{-\tfrac{1}{2}(\frac{x-\theta}{\sigma_1})^2\}$ is clearly bounded we have from (6.25) and the Helly-Bray Theorem that

$$\underset{n\to\infty}{p\lim} \{\Delta_{G_{n+1,n+1}}(x) - \Delta_{G_{n+1}}(x)\} = 0 \qquad \begin{matrix} \text{a.e.}[\mu] \\ \text{a.e.}[Q]. \end{matrix}$$

In order to obtain asymptotic optimality we need to satisfy

$$\sup_n \int_{\Omega} (\theta - \theta_o) dG_n(\theta) < \infty \quad .$$

To do this we assume that for each n, Λ_n is finite a.e.[Q]. Thus applying Theorem 3.1 we have that, for $T = \{t_n\}$ with

$$t_n(x) = \begin{cases} 1 \quad , & \text{if } \Delta_{G_{n+1,n+1}}(x) \geqslant 0 \\ 0 \quad , & \text{if } \Delta_{G_{n+1,n+1}}(x) < 0 \quad , \end{cases}$$

$$R_n(T, G_{n+1}) - R(G_{n+1}) \to 0 \qquad \text{a.e.}[Q].$$

i.e. we have asymptotic optimality in the sense of (6.4). An analagous result holds when considering asymptotic

optimality in the sense of (6.5) since

$$\Lambda_{n+1,n+1} - \Lambda_{n+1} \to 0 \qquad \text{a.e.}[\mu x Q]$$

implies $P^*(\lim_{n\to\infty}\{G_{n+1,n+1}(\theta) - G_{n+1}(\theta)\} = 0$ for all continuity points $\theta) = 1$, where P^* is the probaility with respect to the product measure $\mu x Q$.

A similar result is valid for the selection problem considered in Chapter IV under both the linear loss (4.2) and the zero-one loss (4.14). In fact the procedure $T = \{t_n\}$ where $t_n(\underline{x})$ is defined by (4.48) is asymptotically optimal in the sense of (6.4).

(b) Estimation Under Squared Error Loss

Given $\Lambda_{n+1} = \lambda_{n+1}$, the Bayes procedure for this problem is

$$t_{G_{n+1}}(x) = \frac{\sigma_1^2\lambda_{n+1} + \sigma_2^2 x}{\sigma_1^2 + \sigma_2^2} \quad .$$

Using the estimate $\lambda_{n+1,n+1}$ for λ_{n+1}, define

$$t_n(x) = \frac{\sigma_1^2\lambda_{n+1,n+1} + \sigma_2^2 x}{\sigma_1^2 + \sigma_2^2}$$

For each $s' \in S'$ with $\Lambda_{n+1}(s') = \lambda_{n+1}$ and $\Lambda_{n+1,n+1}(s') = \lambda_{n+1,n+1}$ the excess risk is thus

$$R_n(T,G_{n+1}) - R(G_{n+1}) = E\left|t_n(X) - t_{G_{n+1}}(X)\right|^2$$

$$= \frac{\sigma_i^2}{\sigma^2} \; E|\lambda_{n+1,n+1} - \lambda_{n+1}|^2 \quad .$$

Thus, by Theorem 6.1

$$R_n(T,G_{n+1}) - R(G_{n+1}) \to 0 \qquad \text{a.e.}[Q] \quad .$$

Furthermore since $\text{Var}(X_i|\Lambda_i) = \sigma^2$ a.e.[Q] ,

$$E_Q\left[\frac{1}{(n+1)^2} \sum_{i=1}^{n+1} S_i^2\right] = \frac{1}{(n+1)^2} \sum_{i=1}^{n+1} \sigma^2 = \frac{\sigma^2}{n+1}$$

and we have, under the condition (6.22) that $T = \{t_n\}$ is asymptotically optimal in the sense of (6.5) with

$$E_Q[R_n(T,G_{n+1}) - R(G_{n+1})] = 0((n+1)^{-\delta}).$$

(2) Poisson-Gamma

Suppose $f(x|\theta) = \dfrac{\theta^x e^{-\theta}}{x!}$ $\qquad x = 0,1,2, \ldots ,$

and that, given $\Lambda_i = \lambda_i$, $G_i(\theta)$ has the density

$$g_i(\theta) = \frac{\theta^{\lambda_i - 1}}{\Gamma(\lambda_i)} e^{-\theta} \quad , \qquad \theta > 0.$$

That is, a gamma density with parameters λ_i and 1. For this example it is necessary to assume that the martingale $(\Lambda_i, \mathcal{F}_i)$ is strictly positive in order to avoid degenerate cases. Since $E[X_i|\lambda_i] = \lambda_i$ we use the estimator $\Lambda_{n+1,n+1}$ given by (6.9) and, as for the previous example, (6.25) is satisfied.

(a) The Two-Action Problem

Consider the hypothesis testing problem (3.41) under the linear loss (1.32). Defining $\Delta_{G_{n+1}}(x)$ and $\Delta_{G_{n+1,n+1}}(x)$ similar to that in the previous problem, we need only show that $(\theta_o-\theta)f(x|\theta)$ is continuous and bounded in θ to guarantee asymptotic optimality in the sense of (6.4). In fact

$$(\theta_o-\theta)\frac{\theta^x e^{-\theta}}{x!} = \theta_o f(x|\theta) - (x+1)f(x+1|\theta) ,$$

which is clearly bounded and continuous in θ.

(b) Estimation Under Squared Error Loss

Given $\Lambda_{n+1} = \lambda_{n+1}$ the Bayes procedure is

$$t_{G_{n+1}}(x) = \frac{\lambda_{n+1} + x}{2} .$$

Thus we form the empirical Bayes rule

$$t_{G_{n+1,n+1}}(x) = \frac{\lambda_{n+1,n+1} + x}{2} \quad \text{and}$$

$$R_n(T,G_{n+1}) - R(G_{n+1}) = E\left|\frac{\Lambda_{n+1,n+1} - \Lambda_{n+1}}{2}\right|^2 .$$

For this example $\mathrm{Var}(X_i|\lambda_i) = 2\lambda_i$. Therefore

$$E|\Lambda_{n+1,n+1} - \Lambda_{n+1}|^2 = \frac{2}{(n+1)^2}\sum_{i=1}^{n+1}|\Lambda_i| + \frac{1}{n+1}\sum_{i=1}^{n+1}|\Lambda_i-\Lambda_{n+1}|^2 .$$

Assuming that $\sum_{i=1}^{n+1}\frac{\Lambda_i}{i^2} < \infty$ a.e.[Q] (a condition easily satisfied if Λ_i is bounded a.e.[Q]), then by Theorem 6.1 we have asymptotic optimality for $T = \{t_n\}$ in the sense of (6.4).

If however we assume (6.22), then

$$E_Q[R_n(T,G_{n+1}) - R(G_{n+1})] = E_Q\left[\frac{2}{(n+1)^2}\sum_{i=1}^{n+1}\Lambda_i\right]$$

$$+ E_Q\left[\frac{1}{(n+1)}\sum_{i=1}^{n+1}\Lambda_i-\Lambda_{n+1}\right]^2$$

$$= O((n+1)^{-\delta}),$$

since $E_Q[\Lambda_i] = E_Q[\Lambda_o]$ = constant for each i, because $(\Lambda_i, \mathcal{F}_i)$ is a martingale.

III CONCLUDING REMARKS

The ideas in the empirical Bayes approach form an intuitively attractive line of campaign. Those situations in which one is willing to assume that θ is a random variable but less willing to assume a particular prior distribution for it can be covered by empirical Bayes methods and because the approach makes direct use of previous observations it should offer many advantages over the conventional non-Bayes approach.

The methods developed in this thesis present a partial solution to the empirical Bayes problem where the previous data does not come from identical replications of a fixed decision problem. The examples given are not intended to be exhaustive but are used to illustrate that useful procedures can be obtained from such a situation. The results given here can be extended to the problem of estimating confidence

intervals for the parameter θ. Deely and Zimmer (1969) have studied the normal-normal example and have given confidence intervals for θ, based upon estimation of the mean λ of the fixed a priori distribution $G(\theta)$. These intervals are smaller than confidence intervals based on the maximum likelihood estimate of λ, irrespective of the number of previous observations, $n \geqslant 1$. Thus, in this parametric situation, for the modified problem, we could use the estimator $\lambda_{n+1,n+1}$ for the present mean λ_{n+1}, as the basis for a confidence interval about θ_{n+1}.

In common with much of the empirical Bayes research, it has been assumed here that $f(x|\theta)$ has a known parametric form. Presumably the results here can be extended to cover the non-parametric case (perhaps only with the assumption that $f(x|\theta)$ is a location parameter density) using methods such as those of Johns (1957), Krutchkoff (1967) and Van Ryzin (1970). Their assumption is that at each stage i of observation we have the X_i and in addition we have the collateral information in the form of a supplementary sample W_i (which may be a vector of observations). It is assumed that the components of W_i are mutually independent and independent of X_i and conditionally distributed according to $f(x|\theta)$. All that is now required of this supplementary information is that there exists a measurable function $h(x)$ for which

$$E[h(X)|\theta] = \theta \ .$$

We note that for our location parameter density $h \equiv 1$.

The general problem of determining rates of convergence to optimality for the modified problem, as with that for the usual empirical Bayes problem, remains a very difficult one. The applicability of any empirical Bayes rule depends on the number of previous observations necessary to come within a given distance from the optimal Bayes risk. On one hand if we do not make parametric assumptions on G_n, then, while the asymptotic convergence rates may be quite high, the error estimates for small samples are both very difficult to determine and generally quite inaccurate. On the other hand parametric assumptions, while offering good error rates and easily evaluated bounds, are quite strong restrictions on the problem.

It is difficult to say whether optimality results can be obtained for more general examples in this modified empirical Bayes problem. However with the number of areas into which empirical Bayes techniques are being applied, (see, for example, Krutchkoff (1970)), it is clear that this problem warrants further investigation.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor, Professor John Deely for his assistance and encouragement over the time taken for this study. Thanks must also go to Mrs Janet Warburton for her excellent typing of the manuscript.

REFERENCES

(1) Apostol, T.M., (1974). Mathematical Analysis. Addison-Wesley, Reading, Massachusetts.

(2) Ash, R.B., (1972). Real Analysis and Probability. Academic Press, New York.

(3) Bhattacharya, P.K., (1967). Estimation of a probability density function and its derivatives. Sankhya, Series A, 29, 373-382.

(4) Chung, K.L., (1968). A Course in Probability Theory. Harcourt, New York.

(5) Deely (1965), J.J., (1965). Multiple decision procedures from an empirical Bayes approach. Mimeograph Series No. 45, Department of Statistics, Purdue University.

(6) Deely, J.J. and Kruse, R.L., (1968). Construction of sequences estimating the mixing distribution. Ann. Math. Statist., 39, 286-288.

(7) Deely, J.J. and Zimmer, W.J., (1969). Shorter confidence intervals using prior observations. J. Amer. Statist. Assoc., 64, 378-386.

(8) Deely, J.J. and Zimmer, W.J., (1976). Asymptotic optimality of the empirical Bayes procedure. Ann. Statist., 4, 576-580.

(9) Dunnett, C.W., (1960). On selecting the largest of k normal population means. J. Roy. Statist. Soc., Series B, 22, 1-40.

(10) Ferguson, T.S.,(1967). Mathematical Statistics: a decision theoretic approach. Academic Press, New York.

(11) Johns, M.V., (1957). Nonparametric empirical Bayes procedures. Ann. Math. Statist., 28, 649 - 669.

(12) Johns, M.V. and Van Ryzin, J.R., (1971). Convergence rates for empirical Bayes two-action problems. I. Discrete case. Ann. Math. Statist., 42, 1521-1539.

(13) Johns, M.V. and Van Ryzin, J.R., (1972). Convergence rates for empirical Bayes two-action problems. II. Continuous case. Ann. Math. Statist., 43, 934-947.

(14) Krutchkoff, R.G., (1967). A supplementary sample nonparametric empirical Bayes approach to some statistical decision problems. Biometrika, 54, 451-458.

(15) Krutchkoff, R.G.,(1970). The future of empirical Bayes. From: Proceedings of the Symposium on Empirical Bayes Estimation and Computing in Statistics. Texas Tech. Press, Lubbock, Texas.

(16) Lehman, E.L., (1959). Testing Statistical Hypotheses. J. Wiley and Sons, New York.

(17) Lin, P.E., (1971). Rates of convergence in empirical Bayes estimation problems: Discrete case. Ann. Inst. Statist. Math., 24, 319-325.

(18) Lin, P.E.,(1975). Rates of convergence in empirical Bayes estimation problems: Continuous case. Ann. Statist., 3, 155-164.

(19) Littlewood, B. and Verrall, J.L., (1974). A Bayesian reliability model with a stochastically monotone failure rate. I.E.E.E. Transactions on Reliability, R-23, 108-114.

(20) Loeve, M., (1955). Probability Theory. Van Nostrand Co., New York.

(21) Lukacs, E., (1975). Stochastic Convergence. Academic Press, New York.

(22) Malinvaud, E., (1970). Statistical Methods of Econometrics. North Holland Publishing Co., Amsterdam.

(23) Mara, M.K., (1975). Rates of convergence in empirical Bayes decision problems. M.Sc. Thesis, Univeristy of Canterbury.

(24) Maritz, J.S., (1966). Smooth empirical Bayes estimation for one-parameter discrete distributions. Biometrika, 53, 417-429.

(25) Maritz, J.S., (1967). Smooth empirical Bayes estimation for continuous distributions. Biometrika, 54, 435-450.

(26) Maritz, J.S., (1968). On the smooth empirical Bayes approach to testing of hypotheses and the compound decision problem. Biometrika, 55, 83-100.

(27) Maritz, J.S., (1970). Empirical Bayes Methods. Methuen and Co. Ltd., London.

(28) Meeden, G., (1972). Some admissible empirical Bayes procedures. Ann. Math. Statist., 43, 96-101.

(29) O'Bryan, T. and Susarla, V., (1975). An empirical Bayes estimation problem with non-identical components involving normal distributions. Comm. Statist., 4, 1033-1042.

(30) O'Bryan, T., (1976). Some empirical Bayes results in the case of component problems with varying sample sizes for discrete exponential families. Ann. Statist., 4, 1290-1293.

(31) O'Bryan, T. and Susarla, V., (1976). Rates in the empirical Bayes estimation problem with non-identical components. Case of normal distributions. Ann. Inst. Statist. Math. 28, 389-397.

(32) O'Bryan, T. and Susarla, V., (1977). Empirical Bayes estimation with non-identical components. Continuous case. Austral. J. Statist., 19, 115-125.

(33) Parzen, E., (1962). On estimation of a probability density function and mode. Ann. Math. Statist., 33, 1065-1079.

(34) Raiffa, H. and Schlaifer, R., (1961). Applied Statistical Decision Theory. Harvard University Press, Boston.

(35) Robbins, H., (1955). An empirical Bayes approach to statistics. Proc. Third Berkeley Symposium Math. Statist. and Prob., 1, 157-163.

(36) Robbins, H., (1963). The empirical Bayes approach to testing statistical hypotheses. Rev. Inst. Internat. Statist., 31, 195-208.

(37) Robbins, H., (1964). The empirical Bayes approach to statistical decision problems. Ann. Math. Statist., 35, 1-20.

(38) Rosenblatt, M., (1956). Remarks on some nonparametric estimates of a density function. Ann. Math. Statist., 27, 832-837.

(39) Rutherford, J.R. and Krutchkoff, R.G., (1967). The empirical Bayes approach: estimating the prior distribution. Biometrika, 54, 326-328.

(40) Samuel, E., (1963). An empirical Bayes approach to the testing of certain parametric hypotheses. Ann. Math. Statist., 34, 1370-1385.

(41) Shimi, I.N. and Tsokos, C.P., (1977). The Bayesian and nonparametric approach to reliability studies: A survey of recent work. From: The Theory and Applications of Reliability. Vol 1. ed. by C.P. Tsokos and I.N. Shimi, Academic Press, New York.

(42) Singh, R.S., (1977). Improvement on some known nonparametric uniformly consistent estimators of derivatives of a density. Ann. Statist., 5, 394-399.

(43) Stout, W.F., (1974). Almost Sure Convergence. Academic Press, New York.

(44) Susarla, V. and O'Bryan, T., (1975). An empirical Bayes two action problem with non-identical components for a translated exponential distribution. Comm. Statist., 4, 767-775.

(45) Tainiter, M., (1966). Sequential hypothesis tests for r-dependent marginally stationary processes. Ann. Math. Statist., 37, 90-97.

(46) Van Houwelingen, J.C., (1976). Monotone empirical Bayes tests for the continuous one-parameter exponential family. Ann. Statist., 4, 981-989.

(47) Van Ryzin, J.R., (1970). On some nonparametric empirical Bayes multiple decision problems. From: Nonparametric Techniques in Statistical Inference, ed. by M.L. Puri. Cambridge University Press.

(48) Van Ryzin, J.R. and Susarla, V., (1977). On the empirical Bayes approach to multiple decision problems. Ann. Statist. 5, 172-181.

(49) Wegman, E.J., (1972). Nonparametric probability density estimation; A summary of available methods. Technometrics, 14, 533-546.

(50) Yu, B., (1971). Rates of convergence in empirical Bayes two action and estimation problems and in extended sequence-compound estimation problems. RM-279, Department of Statistics and Probability, Michigan State University.